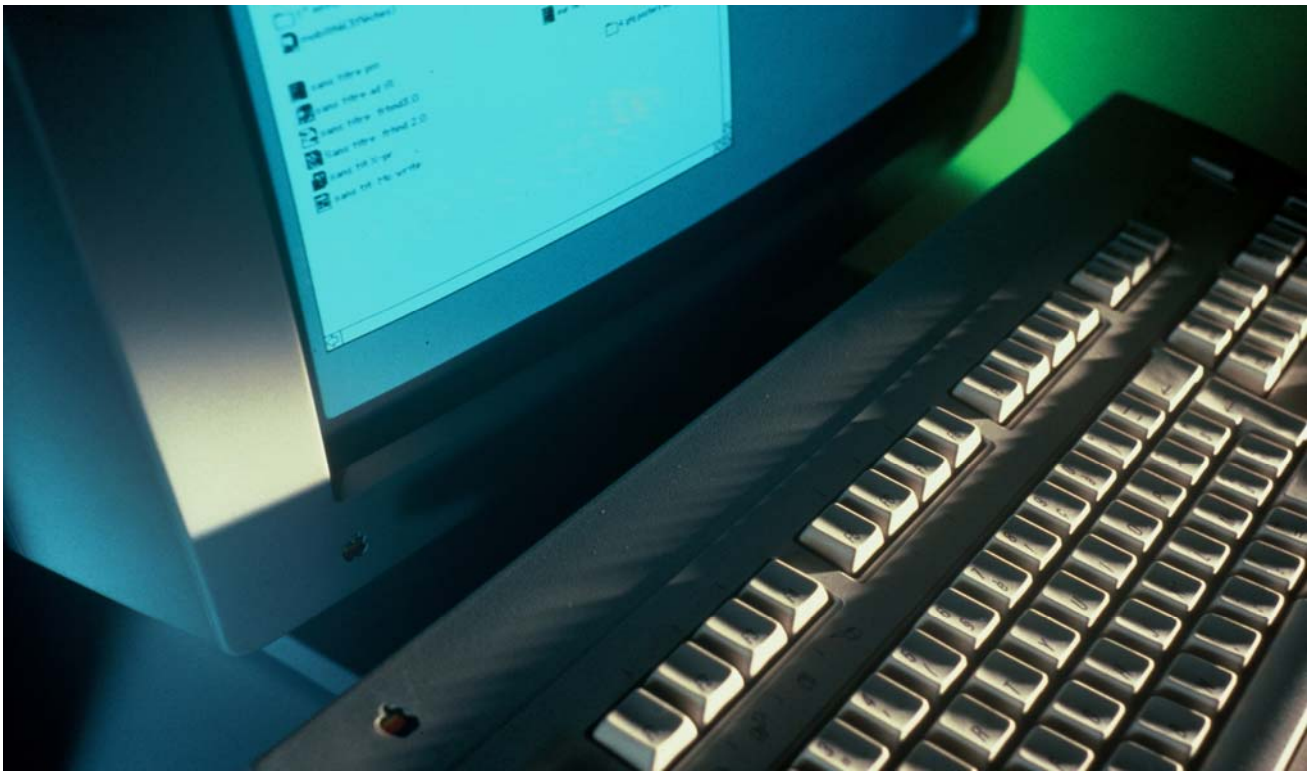
JRC Scientific and Technical Reports

TOWARDS A RESEARCH AGENDA ON COMPUTER-BASED ASSESSMENT

Challenges and needs for European Educational Measurement

Friedrich Scheuermann & Angela Guimarães Pereira (Eds.)



EUR 23306 EN - 2008

JRC EUROPEAN COMMISSION

ipsc Institute for the Protection and Security of the Citizen

qsi

CRELL Centre for Research on Lifelong Learning

The Institute for the Protection and Security of the Citizen provides research-based, systems-oriented support to EU policies so as to protect the citizen against economic and technological risk. The Institute maintains and develops its expertise and networks in information, communication, space and engineering technologies in support of its mission. The strong cross-fertilisation between its nuclear and non-nuclear activities strengthens the expertise it can bring to the benefit of customers in both domains.

European Commission
Joint Research Centre
Institute for the Protection and Security of the Citizen

Contact information
Address: Unit G09, QSI / CRELL
TP-361, Via Enrico Fermi, 2749; 21027 Ispra (VA), Italy
E-mail: angela.pereira@jrc.it, friedrich.scheuermann@jrc.it
Tel.: +39-0332-78.6111
Fax: +39-0332-78.5733

http://ipsc.jrc.ec.europa.eu/
http://www.jrc.ec.europa.eu/

JRC44526

EUR 23306 EN
ISSN 1018-5593

Luxembourg: Office for Official Publications of the European Communities

Printed in Italy

Table of Contents

Introduction

In 2006 the European Parliament and the Council of Europe have passed recommendations on key competences for lifelong learning and the use of a common reference tool to observe and promote progress in terms of the achievement of goals formulated in the "Lisbon strategy" in March 2000 (revised in 2006, see http://ec.europa.eu/growthandjobs/) and its follow-up declarations. For those areas which are not already covered by existing surveys measurements (foreign languages and learning-to-learn skills), indicators for the identification of such skills are needed, as well as effective instruments for carrying out large-scale assessments in Europe. In this context it is hoped that electronic testing could improve the effectiveness of the needed assessments, i.e. to improve identification of skills, by reducing costs of the whole operation (financial efforts, human resources etc.). The European Commission is asked to assist Member States to define the organisational and resource implications for them of the construction and administration of tests, including looking into the possibility of adopting e-testing as the means to administer the tests.

In addition to traditional testing approaches carried out in a paper-pencil mode, there are a variety of aspects needed to be taken into account when computer-based assessment (CBA) is deployed, such as software quality, secure delivery, reliable network (if Internet-based), capacities, support, maintenance, software costs for development and test delivery, including licences.

Any of the delivery modes, whether Paper-Pencil and/or computer-based, comprises advantages and challenges which can hardly be compared, especially in relation to estimated costs. The use of CBA includes additional benefits which can be achieved from an organisational, psychological, analytical and pedagogical perspective. Many experts agree on the overall added value and advantages of e-testing in large scale assessments.

Furthermore, as already pointed out by research presented to PISA Governing Board, October 2006 , change of cultural habits e.g. in terms of reading from computers vs. printed material might suggest an on-going change of assessment forms as well.

Future European surveys will introduce new ways of assessing student achievements. Tests can be calibrated to the specific competence level of each student and become more stimulating, going beyond what can be achieved with traditional multiple choice tests. Simulations provide better means of contextualising skills to real life situations and provide a more complete picture of the actual competence to be assessed.

To date many tools and applications are being developed by commercial enterprises. Commercial packages, underlying organisational concepts, use codes/algorithms usually unpublished which make it difficult to be adopted in different contexts. It is therefore important to take a specific look to what is available on the market as open source software and to reflect about the potential for their implementation in large-scale surveys at a European level.

Overall, CBA is a logic follow-up in the sequence of improvements to be achieved in terms of assessment methodologies, test development, delivery and valorisation of results for multi-purposes. However, a variety of challenges require more research into the barriers posed by the use of technologies, e.g. in terms of computer, performance and security. The Joint Research Centre of the European Commission in Ispra, Italy, is supporting DG Education and Culture in the preparation of future surveys.

The "Quality of Scientific Information" Action (QSI) and the Centre for Research on Lifelong Learning (CRELL) are carrying out a research project on quality criteria of Open Source tools for skills assessment. A workshop was

organised which brought together European key experts from assessment research and practice in order to identify and discuss quality criteria relevant for carrying out large-scale assessments at a European level in terms of objective measurements.

As a consequence, due to the high relevance, participants agreed to engage on further discussion about issues and needs of computer-based assessment in Europe and to publish a report on important aspects to take into account in this field. These activities are seen to be the beginning of the necessary steps to establish a European forum that ensures effective exchange of information and ultimately increases the quality of assessments.

The articles presented here consider several perspectives on computer-based assessment:

Assessment methodologies:

- To what extent does CBA improve methods for skills assessment? Or is this an irreversible trend, due to general pervasion of ICT in everything "we do"? Examples of innovative item types and their potential impact.
- Under which circumstances are further benefits provided in CBA if Computer-Adaptive Testing (CAT) is applied? What are the additional efforts to consider, e.g. in terms of timing, financial resources and delivery, etc.?
- Can large-scale objective measurements (surveys) be linked with supporting the process of teaching and learning?

Implementations / delivery:

- What are the experiences made with large-scale surveys in an international setting? What are the challenges identified, e.g. when internet-based delivery modes are applied?
- What are relevant guidelines and standards to consider? What is there and what is still needed?

Assessment tools:

- Which features should be provided by a software tool, what are the "must" and "nice to have" components of the tools?
- What are the most important quality criteria of software tools for computer-based assessment?
- Relevance of Open Source: Why Open Source? What is the added value Open Source can provide to assessment? How can quality and sustainable benefits be ensured? Examples for implementations, experiences made with large-scale assessments, community support, and security issues.

The articles are framed within a broader discussion on developping a research agenda for European computer-based assessments, highlighting important aspects to take into account. They reflect contributions made during a workshop carried out in November 2007 on "Quality Criteria for Computer-Based Assessment of Skills". The workshop proceedings are accessible online at the CRELL web-site (http://crell.jrc.it/Workshop/200711/cba.htm) and will be published as EU-report in summer 2008.

Friedrich Scheuermann &
Ângela Guimarães Pereira

Ispra, April 2008

New possibilities and challenges for assessment through the use of technology

Romain Martin
University of Luxembourg

The revolution of computer-assisted testing for the domain of educational measurement is an announced revolution that has not really happened. At the end of the eighties, Bunderson, Inouye and Olsen (1989) published their article about the four generations of computerized educational measurement. Thus, it is now two decades ago that these authors saw a continuously growing importance of computerized measurements in the field of education which they described in terms of four different generations of computer-assisted measurement instruments and procedures:

"*It is perhaps inevitable that the recent growth in power and sophistication of computing resources and the widespread dissemination of computers in daily life have brought about irreversible changes in educational measurement.*

Recent developments in computerized measurement are summarized by placing them in a four-generation framework, in which each generation represents a genus of increasing sophistication and power.

Generation 1, Computerized testing (CT): administering conventional tests by computer

Generation 2, Computerized adaptive testing (CAT): tailoring the difficulty or contents of the next piece presented or an aspect of the timing of the next item on the basis of examinees' responses

Generation 3, Continuous measurement (CM): using calibrated measures embedded in a curriculum to continuously and unobtrusively estimate dynamic changes in the student's achievement trajectory and profile as a learner

Generation 4, Intelligent measurement (IM): producing intelligent scoring, interpretation of individual profiles, and advice to learners and teachers, by means of knowledge bases and inferencing procedures." (see abstract of Bunderson et al., 1989)

Bunderson, Inouye and Olsen saw the rapid dissemination of information and communication technologies as one of the major motors for the development and dissemination of computerized educational measurement. It nevertheless turned out that the technological domain is only one facet of this development. The last two decades have indeed known an ever growing spread of information technology which is today very present in the life of every citizen, at least in the industrialized countries. The mobile phones that many pupils use every day to keep in contact with their family and friends would be a medium on which it would at least technically be possible to run computer-assisted tests (see for example Elsmore, Reeves, & Reeves, 2007) and the widespread availability of high-speed internet connections would make a delivery of computer-assisted tests very easy, especially in an educational context. Nevertheless we have to notice that from the four generations of computerized tests, only the first two have been implemented with some success and that presently, we even stick mainly to the first generation, where the main objective is to transpose existing tests on a computer platform.

For the first two generations of tests that rely essentially on the transposition of existing paper and pencil tests, the main advantage of the computer administration compared to paper and pencil relies in the possibility of reliable automatization of data processing procedures that occur in the test administration process: data collection, scoring, reporting etc. The computerized adaptive testing also has its main advantage over paper and pencil in a more efficient administration mode (less items and less testing time), while at the same time keeping measurement precision very high. But also for this second generation measurement models and item formats remain largely identical to those that have been used for paper and pencil testing. The major implementations of the first two generations of computerized educational measurement can thus be found in large scale and often high stakes testing programs, which also exist in paper and pencil formats. These (often commercial) programs benefit most from automatization procedures that permit to increase cost-efficiency. Nevertheless the computerized test administrations have also led to problems especially when computerized

adaptive testing has been used in the context of high stakes testing. The undesirable feature in CATs that item exposure varies greatly with item difficulty and that the most discriminating items are used at high frequencies have indeed been quite problematic for this type of testing in terms of test security (for a more detailed discussion, see Martin, 2003). On the other hand, for large-scale low stakes testing programs like international comparative studies as PISA, computer based administrations have only begun as optional modules in 2006 (with only three participating countries) and will be continued in 2009 by a specific electronic reading assessment that will try to overcome the content domain that can also be covered with paper and pencil instruments. For this type of testing, a sufficient availability of a more or less standardized computer infrastructure in the schools combined with very high demands concerning the standardization of the testing procedure has been one of the major obstacles for the introduction of computer-assisted testing.

In terms of research results, these first two generations of computerized educational measurement have above all produced empirical data on the comparability of paper and pencil administration and computer administration of the same tests, which have generally shown that especially for power tests, this comparability is very high (but with a lesser comparability for speed tests, see Mead & Drasgow, 1993). Other studies have focused on the effects of a potential influence of individual differences in ICT use and familiarity on the results of computer based tests (McDonald, 2002). But the big challenge which has, until now, not really been achieved and intended by generation three and four of the framework described by Bunderson et al. is the use of the new medium in order to go beyond the measurements possible with the paper and pencil format.

As Thornburg (1999) puts it for the field of educational technology
"…no book can contain an interactive multimedia program, and no pencil can be used to build a student's simulation of an ecosystem. The key idea to keep in mind is that the true power of educational technology comes not from replicating things that can be done in other ways, but when it is used to do things that couldn't be done without it" (Thornburg, 1999, p. 7).

An obvious innovative element which can be introduced as by-product of any computer based testing is the collection of user behaviours during test execution. These data might intervene in the scoring procedure, or in the evaluation of the item answer. They can also provide valuable information when analysing collections of answers, for instance to help detecting aberrant response patterns. But the most important potential added-value is dependent on whether these data provide additional useful information on the cognitive functioning or on processing strategies of the subject that are not included in the raw score the subject gets in terms of response correctness. The first candidate for such scrutiny is certainly the response time of subjects which can be obtained very easily even for tests that correspond to transpositions of simple paper and pencil instruments. But a detailed analysis of such response times collected for various computerized paper and pencil tests shows that the interpretation of these response times, while providing interesting information about different processing types, does not permit a univocal interpretation either in terms of general processing speed or in terms of speed-accuracy trade-off mechanisms. An exact interpretation of these response time patterns is merely dependent on the specific task at hand and has to be done on the basis of a detailed analysis regarding the cognitive processes involved (see Martin & Houssemand, 2002 for details).

This fact illustrates very well that what has probably been underestimated in the implementation process of generations three and four described by Bunderson et al. is the fact that besides the technological development, the realization of new methods of computerized educational measurement which seek more dynamic forms of testing relies very much on advances in the field of cognitive science and also of psychometrics. When we try to go beyond the possibilities that were offered by the paper and pencil format it becomes indeed very rapidly obvious that we might need a deeper understanding of the exact processes underlying the tasks under scrutiny in order to make the right interpretations; or that we might need new measurement models in order to take advantage of the new types of data which can be collected with computerized tests. This means that in parallel with the advances made

in terms of technology and of task and data types for which this technology provides access, we also need advances in the fields of cognitive science and psychometrics.

The most obvious potential added value of computer-administered tests lies in the enriched environment that is offered by the computer in terms of display and interaction possibilities. The computer offers indeed the possibility of dynamic displays which are not only limited to static images, but which might include videos, animations, simulations and which offer also the possibility to deliver audio stimuli or even stimuli relating to other sensory modalities, if the corresponding computer-driven feedback devices are available. Also in terms of the modalities of the human-computer interaction and its tracking, computers offer a richer range of possibilities compared to the restriction to hand written feedbacks that are offered by paper and pencil tests. Beyond the reproduction of the paper and pencil response types through the obvious collection of written responses through the computer keyboard and the collection of multiple choice responses through mouse clicks, it is thus imaginable to get feedback through the recording of audio and video stimuli or through the recording of a timed sequence of complex interactions that the subject has with the computer.

A first consequence of the enriched environment offered by the computer might be a more extended use of known experimental paradigms using more dynamic task formats and that provide for example the possibility to evaluate important cognitive constructs such as working memory, attention-related processes etc. These known experimental paradigms have the advantage that the underlying constructs are quite well described by previous research and provide thus a solid theoretical background grounded in the field of cognitive science. On the other hand, they have the disadvantage to address very specific cognitive components that may be constitutive of complex cognitive performances. But while these cognitive components might be well known to cognitive scientists, they may lack face validity in the eyes of stakeholders in the field of education, as they are quite distant from real-world problem solving situations. Major advances in computerized educational measurement have thus to be expected mainly from so-called complex tasks that recreate complex problem solving environments on the computer which are close to real-world problem solving situations (for example in the form of more or less complex simulations). A definition of the characteristics of such complex tasks is provided by Williamson, Bejar and Mislevy (2006, p. 3):

1. "*Completing the task requires the examinee to undergo multiple, non-trivial, domain-relevant steps and/or cognitive processes.*
2. *Multiple elements, or features, of each task performance are captured and considered in the determination of summaries of ability and/or diagnostic feedback.*
3. *There is a high degree of potential variability in the data vectors for each task, reflecting relatively unconstrained work product production.*
4. *The evaluation of the adequacy of task solutions requires the task features to be considered as an independent set, for which assumptions of conditional independence typically do not hold.*"

These complex tasks will also be of major importance for a second strand of potential added value of computerized educational measurement mainly targeted by generations three and four of the Bunderson et al. framework. This major added value was seen in the possibility to integrate learning and testing environments in order to foster directly learning processes by providing continuous formative feedback that can be used either directly by the student or by the teacher in order to organize in a more efficient way the learning environment dedicated to the student. Such an integration of computerized learning and teaching tools will thus generate a demand for diagnostic methods which permit to make qualitative judgments about the current knowledge state of a learner, so that this knowledge state can be taken into account for further learning activities. For such an integration of learning and assessment targeting a continuous evaluation and a possible direct link to pedagogical interventions, it is difficult to imagine that this could be done on the sole basis of classical test theory or even of item response theory which would suppose that one has a sufficient number of calibrated items on every competency dimension targeted by the learning process. Another aspect which does not seem quite satisfactory in these latter measurement models is the conceptualisation of learning progress as a merely quantitative progress on one or many latent traits. For a

future view in which data collected in a computer based learning environment should at the same time provide information about information processing strategies and learning progress of the subject, it seems indeed more promising to view the learning environment as an environment for the processing of complex tasks for which it will be easy to record exact behavioural data due to the presentation of all stimuli in a computer-based system. Instead of imagining the availability of huge item banks of pre-calibrated items for a merely quantitative monitoring of learning progress, it seems then more promising to develop new automated scoring algorithms for such complex computer-based tasks in order to get immediate qualitative and quantitative feedback on learning progress without the necessity of pre-calibrating every task in such a complex problem solving and learning environment. New methodological approaches for the automated scoring of such complex tasks in computer-based testing are currently under development (see Williamson, Mislevy, & Bejar, 2006). Such programs for the integration of learning and testing would also greatly benefit from the availability of open software platforms that would permit a collaborative development of learning and testing instruments and that would provide the possibility to deliver the developed instruments in an efficient way to the learner (see for example Martin, Busana, Latour, & Vandenabeele, 2004).

It can thus be concluded that computer-based assessment instruments are very promising tools for the field of educational measurement. These instruments offer a high potential of added value compared to paper and pencil tests through their data collection and analysis possibilities and through new item formats and test designs taking advantage of the multimedia and interaction facilities offered by computers. But in order to fully benefit from possible added values of computer-administered tests, it will be important to go beyond existing methodological approaches for paper and pencil tests and to provide major research efforts in the upcoming years in the domains of cognitive science and of measurement models.

References

Bunderson, V. C., Inouye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), Educational measurement: Third edition (pp. 367-407). New York: Macmillan.

Elsmore, T. F., Reeves, D. L., & Reeves, A. N. (2007). The ARES(R) test system for palm OS handheld computers. Archives of Clinical Neuropsychology, 22(Supplement 1), 135-144.

Martin, R. (2003). Le testing adaptatif par ordinateur dans la mesure en éducation: potentialités et limites. Psychologie et Psychométrie, 24(2-3), 89-116.

Martin, R., Busana, G., Latour, T., & Vandenabeele, L. (2004). A distributed architecture for internet-based computer-assisted testing. In P. Kommers & G. Richards (Eds.), Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2004 (pp. 114-116). Chesapeake, VA: AACE.

Martin, R., & Houssemand, C. (2002). Intérêts et limites de la chronométrie mentale dans la mesure psychologique. Bulletin de Psychologie, 55(6), 605-614.

McDonald, A. S. (2002). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. Computers & Education, 39(3), 299-312.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. Psychological Bulletin, 114(3), 449-458.

Thornburg, D. D. (1999). Technology in K-12 Education: Envisioning a New Future. Retrieved March 13th, 2008, from http://www.eric.ed.gov/

Williamson, D. M., Bejar, I. I., & Mislevy, R. J. (2006). Automated scoring of complex tasks in computer-based testing: An introduction. In D. M. Williamson, R. J. Mislevy & I. I. Bejar (Eds.), Automated scoring of complex tasks in computer-based testing (pp. 1-13). Mahwah, N.J.: Lawrence Erlbaum Associates.

Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (2006). Automated scoring of complex tasks in computer-based testing. Mahwah, N.J.: Lawrence Erlbaum Associates.

The author:
Romain Martin
University of Luxembourg
EMACS research unit (Educational Measurement and Applied Cognitive Science)
FLSHASE-Campus Walferdange, B.P.2
L-7201 Walferdange
E-mail: romain.martin@uni.lu

Romain Martin is Professor of Psychology and Educational Science at the University of Luxembourg and head of the EMACS research unit. He has initiated the TAO project which seeks the development of a collaborative open-source platform for computer-based assessment. He is currently responsible for a number of research projects developing new computer-aided assessment tools for the field of education. He is also actively involved in projects seeking to introduce computer-based assessment in school monitoring and large-scale assessment programs.

Changing Icelandic national testing from traditional paper and pencil based tests to computer based assessment: Some background, challenges and problems to overcome

Júlíus K. Björnsson
Educational Testing Institute

Summary
The current Icelandic national testing system is shortly described and imminent changes outlined. These changes however entail a number of problems, especially in ensuring that a continuity in testing is maintained, that the same competencies are assessed and that systematic differences between paper and pencil testing and computer - based assessment do not lead to erroneous conclusions about changes in proficiency or other systemetic changes attributable to the mode of testing. Some relevant results from the PISA 2006 Computer Based Assessment of Science (CBAS) are presented, showing that there may be systematic differences between countries in how strongly the mode of test administration influences test results. Finally, some conclusions and directions for further study are presented.

Introduction

Iceland has a long tradition of national testing, the first tests being held in 1929 for twelve year old children which at that time were at the end of compulsory schooling. The system was changed a number of times during the twentieth century as the educational system evolved. Iceland basically inherited a Danish educational system in 1918 when the country became independent and traditionally the state has been responsible for all education. This has been, however, gradually changing and the local communities took over responsibility for primary and lower secondary schools in 1996 and a number of private schools and universities have been established recently. The state is, however, responsible for delivering a central curriculum for primary and secondary schools, which they are obligated to follow, although they all have a certain level of freedom within that framework.

Teacher training has also been changing gradually, in 1975 it was moved up to university level from being in special teacher colleges, and now the intention is to increase the training requirements so that all teachers of primary and lower secondary schools get five years of training and an MA degree.

Approximately every 10-15 years the national testing program has changed along with the relevant laws about compulsory schooling and the last big change was in 1993 when modern psychometric methods were introduced to the testing construction and since then the national testing system has evolved gradually. Now it is changing again, however, as described in the following.

The current testing system consists of tests in Icelandic (reading writing etc) and mathematics for both the 4th and 7th grades (10 and 12 year olds) and for the tenth grade (15 year olds) there are tests in Icelandic, mathematics, English, Danish, social studies and science. The tests for the 4th and 7th grades are obligatory for all pupils and their purpose if first and foremost to check on individual progress and to gather information about the performance of the whole system. As stated in the relevant regulation published by the Ministry of Education the purpose of the tests is:

- To check that both goals and subgoals in the national curriculum in each subject have been reached (or how many pupils have reached them).
- Give teachers directions for the continued education of each student.
- Gather information about schools, how they do in each subject compared to other schools.

And additionally in the tenth grade:

- Collect information for the upper secondary schools about each students standing. (i.e. produce intake information for the upper secondary schools).

The purpose of the whole system is, therefore, to gather information about the whole system, about each region in the country, each school district, each school, each class and the individual student. There are thus multiple purposes behind the tests and this entails considerable psychometric challenges as it is

admittedly very difficult to construct single tests that fullfill all of these goals simultaneously in an adequate manner in the same test.

All the current tests are written and constructed at the Educational Testing Institute (ETI) in cooperation with a large group of teachers and experts in each subject. The tests are piloted in relevant groups of students and constructed in such a way that they are psychometrically sound, with adequate reliability and validity. The general rule for test administration is that every test is held at the same time in all schools. This goes for all the national tests for the 4th, 7th and 10 grades. Students in the 10th grade can choose how many of the tests they take, i.e. if they want to enter upper secondary school. In practice most of them take Icelandic, math and English.

All tests are centrally graded and scored at the ETI in order to ensure scorer reliability for open-ended questions and in order to ensure coordinated grades for everyone. The tests for the 4th and 7th grades are held every year in October and tests for the 10th grade every year at the beginning for May. Students from the 8th and 9th grades can take the 10th grade tests is they so choose.

Every student gets a grade 3-4 weeks after taking the test and various report are produced for schools, school districts and for the educational system as a whole.

System changes

The whole school system has recently been more and more oriented towards individualized teaching and learning. Almost everyone agrees that these changes are very desirable and should be very beneficial for students, but in practice classes are just as large as they have been for many years and the teacher-student ratio is relatively unchanged. The national tests have in recent years been criticized for being restrictive, for not being able to test all aspects of the students learning and perhaps primarily for being very controlling for both teachers and students work. There are reported instances of massive drilling for the tests, where students are drilled on old tests for a considerable period of time before taking the tests themselves. Although research has shown this to be counterproductive, it still goes on to some extent.

The ETI has therefore recently proposed to change the testing system over to computerized adaptive testing and this proposal has been enthusiastically received by politicians, schools, teachers and parents.

Computerized testing

With a new testing system based on Computer Based Assessment methods, and especially adaptive testing, the following advantages are possible:

- Shorter testing time for each student.
- A better student-test fit with adaptive tests.
- Quicker results for each student than is now possible.
- Much higher precision in the measurement, especially at high and low achievement levels, ie a more equiprecise test.
- A more enjoyable and better testing experience for the students.
- Less stress and press on all concerned as the tests will not take place all at the same time but will be administered over a period of time each year.
- Testing with the medium (computers) that all students are basing more and more of their learning on.
- The possibility of rich items and multimedia content.
- The reuse of test items, which in Iceland has been impossible until now as it has been obligatory to publish all tests after they have been held.
- Cheaper and quicker coding of test responses.
- Better information about the student group, the schools, school districts and the whole educational system.
- Trend information which has not been possible to get because everything gets published, but with the reuse of items this possibility opens up a new way of looking at changes in achievement over time.

Even though the above mentioned advantages are very appealing to everyone concerned, there are, of course, costs and efforts required when introducing a new system of electronic testing.

The startup of such a system is expensive - especially the development of an item bank for adaptive testing purposes - but this is a necessary expense in order to be able to start the system. It is also clear that some competencies cannot be tested, but in reality this is not so different from the situation with

paper and pencil tests, as that testing mode also has its limitations of course. The requirements for both technical and psychometric competences are higher than for the older type of tests, but those requirements are more or less known and therefore in themselves not a problem. However, testing in all schools will require considerable resources, both technical ones and specially trained personell to oversee and administer the new type of tests.

But some central questions remain when changing from a traditional paper and pencil testing system over to CBA with adaptive testing. Some of those are the following:

- Are the same competencies assessed in a paper and pencil and electronic version of a test of the same subject?
- If the tests are measuring the same competency, are there other differences, such as gender differences, differences related to computer competencies or any other systematic differences?
- Do some students get unfair advantages/disadvantages when tested with CBA?
- When changing testing mode/method is it not necessary to know what the differences are in order to be able to compensate for them so that test results are comparable over time?

Some results from the PISA Computer Based Assessment of Science (CBAS)

In order to begin to answer some of the questions posed above, some results from the PISA CBAS will be shown here. It is admittedly difficult to compare the paper and pencil PISA test of science and the CBAS as these two tests of science did use different items in addition to being administered in a completely different way. It was clear from the outset that the results and especially a comparison of the results from these two tests would have to be done very carefully as the differences were considerable, especially concerning a few important variables. Reading load (the amount of text) was much less in the CBAS than in the paper and pencil version and it is already known that the correlation between reading and science on the PISA is very high, above 0,6. Therefore it was very probable that students with lower reading proficiency would do relatively better on the CBAS than on the PISA 2006 paper test.

The CBAS was done in three countries, Iceland, Denmark and Korea and was administered to a subsample of the students participating in the PISA 2006 assessment. After these students had taken the conventional test, they took the CBAS, either the same day or very shortly thereafter. The administration was heavily standardized, all testing employed the same type of computer, a standardized laptop, the same software of course and similar testing environment. All data were rescaled for the three countries involved and therefore the scores on the CBAS are not comparable to the PISA 2006 scores themselves. The date were rescaled to a scale with a mean of 500 and a standard deviation of 100 and this means that not only the performance means for the countries are different from those obtained in the PISA 2006, but also the variation of the whole scale.

Average scores for the three countries participating in the CBAS are shown in table 1.

	PISA 2006			**CBAS 2006**		
	Female	Male	Total	Female	Male	**Total**
Denmark	469 (8,2)	492 (6,9)	480 (6,2)	440 (7,4)	485 (6,2)	**462 (5,3)**
Iceland	474 (2,5)	467 (2,6)	470 (1,7)	459 (2,2)	484 (2,5)	**471 (1,6)**
Korea	**502 (6,4)**	**501 (6,1)**	**502 (4,3)**	**489 (7,2)**	**515 (6,5)**	**503 (4,8)**

Table 1. Scores from the three countries

The table shows that the results change considerable according to mode of testing, and the gender difference changes are especially notable. In Iceland and to some extent in Korea they change not only in magnitide but also in direction. The females in Iceland are considerable better on the PISA 2006 test but this is turned around in the CBAS with the boys performing considerably better. The same thing happens really in Denmark but in such a way that the gap between boys and girls which is considerable in the PISA 2006 widens markedly in the CBAS, becoming almost half a standard deviation. The same thing happens in Korea where there is no gender difference in the PISA 2006 but a difference of 25 points in favor of the boys on the CBAS. A strong correlation exists between the results obtained in the three countries between scores on the CBAS and all three domains tested in the PISA 2006, and these are shown in table 2.

	Science PISA 2006		Reading PISA 2006		Math PISA 2006	
	F	M	F	M	F	M
Denmark	0,89	0,90	0,81	0,77	0,83	**0,84**
Iceland	0,78	0,79	0,71	0,73	0,73	**0,76**
Korea	**0,88**	**0,89**	**0,77**	**0,77**	**0,86**	**0,87**

Table 2. Correlations between test methods

When examining these correlations it is notable that they are fairly similar across all three countries, the correlation with PISA 2006 science is almost 0,9 in both Denmark and Korea but considerably lower in Iceland in both genders. Correlations between the CBAS and PISA 2006 reading is roughly comparable across all countries but again Iceland is a little bit different from the other two in Mathematics having again a lower correlation with PISA 2006 math than both Korea and Denmark.

It is, of course, difficult to interpret these scores across the two studies as the means are not directly comparable, but the numbers appear to indicate that not only do the boys do much better on a computerized test, most probably the girls are doing worse on the same test than on the paper and pencil test. Therefore these results can potentially have a great effect, but it is clear that these differences must be much better studied; the same type of items and content must be assessed with both modes of administration in order to make us able to conclude anything about the differences between these types of tests. Furthermore, it is probable based on the differences between these three countries in the correlations with PISA 2006 results that the same relationships do not hold across the countries. The correlations tend to be considerably lower in Iceland, especially with science and Math but they are approximately the same with PISA 2006 reading, probably reflecting the reduced reading load in the CBAS.

We cannot fully explain yet the differences shown above between these three countries, but in order to show further the complexity of this situation, what follows is some data from the questionnaire that all CBAS students answered after taking the computerised test. This was a short questionnaire about attitudes to conventional test and CBA tests, and questions about which mode of testing the students would prefer. Table 3 presents the answers on whether the students found the CBAS an enjoyable experience.

	Agree strongly		Agree		Disagree		Disagree strongly	
	F	M	F	M	F	M	F	**M**
Denmark	29,6	35,7	57,2	49,1	8,7	8,7	4,5	6,5
Iceland	4,6	9,4	46,6	40,4	33,6	27,3	15,2	22,9
Korea	30,1	34,6	55,2	48,8	12,0	11,2	2,7	5,4

Table 3. Proportions of students endorsing different options about the statement: "I found the computerised test enjoyable."

Here we see the same pattern of answers in Korea and Denmark but the Icelandic students differ markedly as very few of them are agreeing with the statement strongly and almost a quarter of them disagreeing strongly. If this attitude towards the test method has any effect on the results, then this could perhaps explain some of the differences observed earlier between Iceland on one hand and Denmark and Korea on the other. There is also a much greater gender difference in Iceland than in the other countries with boys enjoying the computerised test more than girls. Here the differences between genders are much larger in Iceland than in the other countries although the tendency is the same. Table 4 is about the same thing but from the other perspective.

	Agree strongly		Agree		Disagree		Disagree strongly	
	F	M	F	M	F	M	F	**M**
Denmark	4,3	4,6	40,1	31,6	41,1	38,7	14,6	**25,2**
Iceland	2,8	2,7	21,7	19,4	39,7	39,1	35,8	**38,8**
Korea	**6,8**	**8,0**	**55,2**	**48,8**	**12,0**	**11,2**	**2,7**	**5,4**

Table 4. Proportions of students endorsing different options about the statement: "I found the paper and pencil test enjoyable."

Here we see some differences between all three countries, although the common thing appears to be that very few students in all countries say that they agree strongly with the statement. Nobody likes a test! However, when examining the other end of the answer spectrum Korea shows the exception with very few students disagreeing strongly with the statement. The strongest dislike appears to be in Iceland with Denmark a bit behind and very few students in Korea that disagree strongly with the statement that they enjoyed the paper and pencil test.

So once again, as has been done so many times in the literature, it has been demonstrated here that, if someone says that they dislike something, this does not necessarily mean that the same person likes the phenomenon. Very few Koreans dislike the PP test and very few like it. So here, there are again differences between these three countries that have the potential of influencing the results on the test, but the interrelationships between these attitudes and the performance on both tests remain to be explored.

Finally, results from asking the students to compare the two modes of administering a science test are shown in table 5.

	Two hour PP test and nothing with a computer.		A test which is one hour PP and one hour Computer		Two hour test on computer	
	F	M	F	M	F	M
Denmark	7,1	4,2	39,6	37,8	53,3	58,0
Iceland	11,9	18,6	42,8	32,8	45,3	48,6
Korea	5,7	8,0	42,2	30,3	52,1	61,7

Table 5. Proportions of students choosing between three modes of test administration.

The table indicates that around half of the CBAS students would prefer a completely computerized test and suprisingly Iceland had the highest number of students who perferred a paper and pencil test, more than twice the number in Korea and again surprisingly, of those preferring a PP test the boys were more numerous conflicting with the fact that they appear to do relatively much better on a CBA test. So perhaps they do not know to well what is good for them. But again, these attitudes have to be related to performance in order to understand these rather strange proportions better.

Conclusions

It appears probable, from this admittedly short and superficial analysis, that the PISA 2006 and CBAS were measuring the same competencies. However, it emerges also that the gender differences are variable across countries, that attitudes towards computerized testing are different in different countries and not necessarily a polarization of either liking paper and pencil tests or computerized tests. The picture is probably much more complicated and the relationships between the performance on the CBAS and the PISA 2006 with attitudes toward the different modes of testing will be explored in a further publication. Additionally there may be cultural factors which potentially could explain some of the differences observed here in attitudes to these modes of testing.

The perhaps most important conclusion one can draw from these data is that general principles about the differences between traditional paper and pencil testing and CBA are very probably not the same in different countries. Therefore, it would be unadvisable to assume that the same differences apply everywhere and this further underlines the fact that a country which wishes to change from traditional testing methods to the new CBA methods must do so very carefully and along the way test meticulously which differences appear to be dominant in the country. It has therefore been decided that in Iceland the change from the old testing system over to the new one will be done via a controlled experiment where it will be possible to examine the above discussed differences as meticulously as possible, comparing traditional paper and pencil testing with both linear CBA and adaptive testing. But this is another discussion and will be the subject matter of future research.

References:

OECD (2007). PISA 2006: Science Competencies for Tomorrow's World. OECD: Paris.
Almar M. Halldórson, Ragnar F. Ólafsson & Júlíus K. Björnsson. Færni og þekking nemenda við lok grunnskóla: Helstu niðurstöður PISA 2006 í náttúrufræði, stærðfræði og lesskilningi. Námsmatsstofnun 2007. (Icelandic National PISA 2006 report)

The author:

Júlíus K. Björnsson
Educational Testing Institute
Borgartúni 7a
105 Reykjavík, Iceland
E-Mail: julkb@namsmat.is

Julius K. Björnsson: Currently director of the Icelandic Educational testing Institute. Training in psychology, graduate in Clinical Psychology, lecturer in psychometrics and psychophysiology at the University of Iceland for ten years.

Accelerating the assessment agenda: thinking outside the black box

Denise Whitelock
The Open University

Abstract:
Over the last 10 years, learning and teaching in higher education have benefited from advances in social constructivist and situated learning research (Laurillard, 1993). In contrast, assessment has remained largely transmission orientated in both conception and in practice (see Knight & Yorke, 2003). This paper examines a number of recent developments, which exhibit innovation in electronic assessment developed at the UK's Open University. This paper argues for the development of new forms of e-assessment where the main driver is that of sound pedagogy rather than state of the art technological know-how and where open source products can move the field forward.

Introduction

As teaching and learning cannot be separated from each other in practice it is difficult to think about learning without including assessment. It is well documented that assessment drives learning (see Rowntree, 1977) and teachers too, especially in the UK, are acutely aware of assessment targets with the introduction of league tables, (see the UK's Department for Children, Schools & Families Achievement and Attainment tables **http://www.dcsf.gov.uk/performancetables**). Other types of testing such as the Programme for International Students Assessment (PISA) (**http://www.pisa.oecd.org/pages/**) and Trends in International Mathematics and Science Study (TIMSS) (**http://nces.ed.gov/timss**) provide information to the bigger league table of the European Union. Although they are laudable how can the latter, with their well constructed tests, assist the students learning, profit the teaching and move us forward along the assessment agenda? By this I mean constructing a creative and collaborative milieu where the climate is not one of 'teaching for the assessment', but rather more of 'assessment for learning'. This paper argues for the development of new forms of e-assessment where the main driver is that of sound pedagogy rather than state of the art technological know-how and where open source products can move the field forward.

The constructivist learning push

Over the last 10 years, learning and teaching in higher education have benefited from advances in social constructivist and situated learning research (Laurillard, 1993). In contrast, assessment has remained largely transmission orientated in both conception and in practice (see Knight & Yorke, 2003). This is especially true in higher education where the teachers' role is usually to judge student work and to deliver feedback (as comments or marks) rather than to involve students as active participants in assessment processes.

However, recent research as well as highlighting the problems also holds the key to unlocking the assessment logjam. Firstly, there is recognition that the role of the student in assessment processes has until now been under-theorised and that this has made it difficult to address the relevant issues effectively. Students do not learn through passive receipt of teacher-delivered feedback. Rather, research shows that effective learning requires that students actively decode feedback information, internalise it and use it to make judgements of their own work (Boud, 2000; Gardner, 2006; Sadler, 1989). This, and other findings, emphasise that learners engage in the same assessment acts as their teachers and that self-assessment is integral to the students use of feedback information. Indeed, Nicol and Macfarlane-Dick (2006) argue that formative assessment processes should actually be designed to 'empower students as self-regulated learners'.

Another recent research direction has been to develop broader theoretical foundation for learning and assessment practice. The Assessment Reform Group (Gardner, 2006) have begun work on a theory of assessment relevant to the school classroom (Black & Wiliam, 2006). They adopt a community of practice approach (Lave & Wenger, 1991) and interpret the interactions of assessment tools, subjects and outcomes from the perspective of activity theory (Kutti, 1996). There are four key

components within this framework: (i) teachers, learners and the subject discipline (ii) the teacher's role and the regulation of learning (iii) feedback and the student-teacher interaction and (iv) the teacher's role in learning. Black and Wiliam (2006) argue that one function of this framework will be 'to guide the optimum choice of strategies to improve pedagogy'. Other researchers who have identified the need for a more complete development of theory in order to enhance pedagogic practice are Yorke (2003) and James (2006).

Another area of research is that showing the critical effects of socio-emotional factors in the design of assessment. Dweck and her colleagues (Dweck, 1999; Dweck, Mangels, & Good, 2004) have shown that cognitive benefits in assessment are highly dependent on emotional and motivational factor: beliefs and goals affect basic attentional and cognitive processes. In particular, this research shows that even small interventions in assessment practice can have dramatic impacts on learning processes and outcomes: e.g. focusing students on learning goals rather than performance goals before task engagement, praising effort rather than intellectual ability.

The vision for e-assessment in 2014 which is documented in Whitelock and Brasher's (2006) Roadmap study reveals that experts called for a pedagogically driven model rather than a technologically and standards led framework to lead future developments in this area. Experts believed that students will take more control of their own learning and become more reflective.

The future would be one of more 'on-demand testing' that will assist students to realise their own potential and e-portfolios will help them to present themselves and their work in a more personalised manner. This notion is also supported by the then DfES (Department for Education and Skills) agenda to promote "personalised" learning, with e- assessment playing a large role. However the production of such software is costly and requires large multidisciplined teams. One of the ways forward then is to adopt the open source model as advocated by the JISC and the UK's Open University, which has funded many successful in house developments as illustrated below has adopted Moodle, an open source application as its VLE.

The role of feedback in assessment

One of the challenges for e-assessment and of today's education is that students are expecting better feedback, more frequently, and more quickly. Unfortunately, in today's educational climate, the resource pressures are higher, and feedback is often produced under greater time pressure, and often later.

This raises the question of what is meant by *feedback*? The way our team (Watt et al, 2006) have defined feedback is that it is seen as additional tutoring that is tailored to the learner's current needs. In the simplest case, this means that there is a mismatch between students' and the tutors' conceptual models and the feedback is reducing or correcting this mismatch, very much as feedback is used in cybernetic systems. This is not an accident, for the cybernetic analogy was based on Pask's (1976) work, which has been a strong influence on practice in this area (e.g., Laurillard, 1993).

The Open University has been building feedback systems over a number of years. Computer marked assignments consisting of a series of multiple questions together with tutor marked assignments have provided the core of assessment for our courses for a number of years. There is now a move, like the school examination boards, towards synchronous electronic examinations. A study was undertaken by Thomas et al (2002) who found that post graduate computer students who completed a synchronous examination in their own home were not deterred by it and were happy to sit further examinations in this manner.

Another course at the Open University i.e. 'Maths for Science' aimed to take the findings of Thomas et al's study one step further. It not only offered students a web-based examination in their own home but also provided them with immediate feedback and assistance when they submitted their individual answers to each question. This design drew on the findings from the interactive self-assessment questions initially devised for an undergraduate science course 'Discovering Science' (Whitelock, 1999) which offered different levels of feedback when the student failed to answer a question correctly and a similar system has also been employed by Pitcher et al (2002).

The Maths for Science software was built to deduct marks according to the amount of feedback given to a student when they answered a question. It was anticipated that the provision of partial marks for second and third attempts would encourage students to try questions that they might otherwise have ignored through lack of confidence or incomplete knowledge. Again, at its simplest the system awarded 100% of the marks for a question answered correctly at the first attempt, 65% to students who answered correctly after they received a text hint to help them select the correct response and 35% to students who gave the correct answer after receiving two sets of text hints. All students received a final text message, which explained the correct solution to the question, which had just been answered. This type of feedback is relevant to both student learning and the grading process. It integrates assessment into the teaching and learning feedback loop, and introduces a new level of discourse into the teaching cycle as advocated by Laurillard, (1993).

'Maths for Science' was a short course (worth 10 credits only) and was designed to teach students the necessary algebraic skills to progress to second level scientific courses. The maintenance of short courses is a resource heavy exercise, and online delivery reduced the amount of time required to process results and awards. Unlike long Open University courses (60 credits), short courses were produced for students to enhance their own study skills, and therefore little benefit would be gained from cheating in the examinations. All the students managed to take the examination at home after a practice examination was attempted. They found it easy to use and felt they learnt a lot with this format, especially when the reasoning for each correct solution was revealed (Whitelock and Raw 2003). They were also pleased to obtain partial credit for their answers.
Other systems have shown the benefits of providing minimal immediate feedback to students for university examinations taken not at home but in a room full of colleagues working with computers under normal examination conditions. This modus operandi has been adopted by the Geology department at Derby University who developed TRIADS software which has been used for end of year examination (**http://www.derby.ac.uk/assess/newdemo/mainmenu.html**).

The above examples all suggest that providing feedback during electronic assessment has a broad appeal for students. It has also been documented that this type of feedback enhances learning in a variety of fields (Elliott, 1998; Phye and Bender, 1989; Brosvic et al 1997). A delay on the other hand may reduce the effectiveness of feedback (Gaynor 1981; Gibbs and Simpson, 2004). These findings indicate that systems, which provide immediate feedback, have clear advantages for students engaging in a learning dialogue during and after electronic assessment is of value but how can students collaborate on electronic assignments?

This notion that knowledge and understanding are constituted in and through interaction has considerable currency and a growing body of work emphasises the need to understand the dynamic processes involved in the joint creation of meaning, knowledge and understanding (e.g. Grossen & Bachmann, 2000; Murphy, 2000; Littleton, Miell & Faulkner, 2004; Miell & Littleton, 2004). The theoretical background here is of social constructivism which builds upon the notion of interaction with significant others in the learning process. Creating a sense of presence online and an environment that can be used to encourage students to work collaboratively on interactive assessment tasks is certainly a challenge.

Our most recent project has embellished an application known as "BuddySpace" (see Vogiazou et al, 2005), which was developed by KMi at the Open University to provide a large–scale informal environment for collaborative work, learning and play. It utilises the findings from distance education practice (Whitelock et al, 2000) that the presence of peer-group members can enhance the emotional well-being of isolated learners and improve problem-solving performance and learning. Rheingold (2002) too discusses the power of social cohesiveness that can be achieved through the simple knowledge of the presence and location of others in both virtual and real spaces.

BuddySpace builds on the notion of an Instant Messaging system that has a distinct form of user visualisation that is superior to a conventional 'buddy list'. In fact, BuddySpace provides maps to represent each group member's location (see Figure 1 below).

Figure 1: BuddySpace location map

This allows a new member of the group to see if there are any other members from the same course living close by. BuddySpace is a piece of open-source software and, to date; Eisenstadt reports that it has been downloaded by some 19,000 users. Presence and availability can also be conveyed with this system showing 'available for chat', 'do not disturb'; 'low attention' or 'online but elsewhere'.

In order to give students the opportunity to work together on complex formative assessment tasks we added other features to BuddySpace. These features allow users to add details of their expertise and interests into a database so that other users could find them in order to seek out their expertise on a variety of topics and to 'yolk' PCs together so that two students could see and synchronously interact with a software simulation. Hence BuddyFinder and SIMLINK were developed by IET and Kmi.

In Figure 2 below, the two 'students', Chris and Simon, both see the same set of sliders and graphs on their screens. As one student moves a slider, the other student sees the same action on his screen. In other words both students view identical screens at the same time. An action on one student's screen is mirrored on the others. (The simulation shown in Figure 2 is a version of the Global Warming simulation used on the science foundation course)

The goals of this particular work is to build open source applications that will assist science and technology courses to construct complex problem solving activities that require a partner to assist with their solution as well as more straightforward feedback systems for individuals to use to test their understanding of a particular domain.

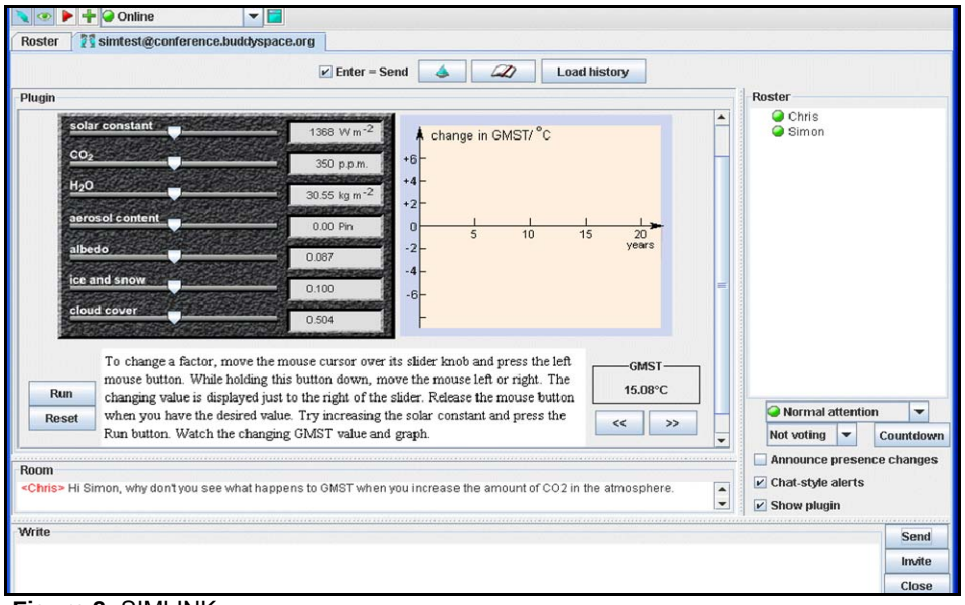


Figure 2: SIMLINK

Because feedback is very much at the cutting edge of personal learning, Whitelock and Watts (2007) wanted to see how we could work with tutors to improve the quality of their feedback. To achieve this, we have been working on tools to provide tutors with opportunities to reflect on their feedback. The latest of these, Open Mentor, (http://kn.open.ac.uk/workspace.cfm?wpid=4126) is an open source tool which tutors can use to analyse, visualise, and compare their use of feedback. For this application feedback was considered not as error correction, but as part of the dialogue between student and tutor. This is important for several reasons: first, thinking of students as making errors is unhelpful – as Norman (1988) says, errors are better thought of as approximations to correct action.

Thinking of the student as making mistakes may lead to a more negative perception of their behaviour than is appropriate. Secondly, learners actually need to test out the boundaries of their knowledge in a safe environment, where their predictions may not be correct, without expecting to be penalised for it. Finally, feedback does not really imply guidance (i.e. planning for the future) and we wanted to incorporate that type of support without resorting to the rather clunky 'feed-forward'.

The lessons learned from Open Mentor can be applied to feedback to students during or immediately after electronic assessments. This will assist them to take more control of their own learning and will also recognise their anxiety which is provoked by the test environment. This is a position argued by McKillop (2004) after she asked students to tell stories about their assessment experiences in an on-line, blog-style environment. This constructivist approach also aimed to involve students in reflective and collaborative experiences of their assessment experiences.

The insights gained from this project are currently being applied to a new feedback system developed at the Open University for electronic formative assessment of history students that uses free text entry with automatic marking and is known as Open Comment. (http://kn.open.ac.uk/workspace.cfm?wpid=8236)

Conclusions

In today's educational climate, with the continued pressure on staff resources, making individual learning work is always going to be a challenge. Assessment is the main keystone to learning and lack of submission of assessments often leads to student drop out in higher education (Simpson, 2003). However it is achievable, so long as we manage to maintain our empathy with the learner. Embracing constructivism and developing new types of e-assessment tools can help us achieve this by giving us frameworks where we can reflect on our social interaction, and ensure that it provides the emotional support as well as the conceptual guidance that our learners need.

Technology to enhance assessment is still in its early days, but the problems are not technical: assessment raises far wider social issues, and technologists have struggled in the past to resolve these issues with the respect they deserve. A community of open source developers collaborating on these big issues can offer a new way forward to these challenges. e-Assessment is starting to deliver potential improvements; but there is still much work to be done.

Acknowledgements

The author would like to thank all her colleagues at the Open University who have worked on the various projects mentioned in this paper. She is indebted to them for their contributions and also special thanks are due to Stuart Watt for his insightful involvement and good humour.

References

Black, P., & Wiliam, D. (2006). Developing a theory of formative assessment. In J. Gardner (Ed.), Assessment and Learning (pp. 81-100). London: Sage Publications.

Boud, D. (2000). Sustainable assessment: rethinking assessment for the learning society. Studies in Continuing Education,, 22(2), 151-167

Brosvic, G.M., Walker, M.A., Perry, N., Degnan, S. and Dihoff, R.E. (1997) 'Illusion decrement as a function of duration of inspection and figure type', Perceptual and Motor Skills, Vol. 84, pp.779–783

DFES (2005) The e-Strategy - Harnessing Technology: Transforming learning and children's services. Available from: www.dfes.gov.uk/publications/e-strategy .

Dweck, C. S. (1999). Self-Theories: Their role in motivation, personality and development. Philadelphia: Psychology Press.

Dweck, C. S., Mangels, J. A., & Good, C. (2004). Motivational effects on attention, cognition and performance. In D. Y. Dai & R. J. Sternberg (Eds.), Motivation, emotion, and cognition: Integrated perspectives on intellectual functioning: Lawrence Erlbaum Associates.

Elliott, D. (1998) 'The influence of visual target and limb information on manual aiming', Canadian Journal of Psychology, Vol. 42, pp.57–68.

Gardner, J. (Ed.). (2006). Assessment and Learning. London: Sage Publications.

Gaynor, P. (1981) 'The effect of feedback delay on retention of computer-based mathematical material', Journal of Computer-Based Instruction, Vol. 8, pp.28–34.

Gibbs, G., & Simpson, C. (2004). Conditions under which assessment supports students' learning. Learning and Teaching in Higher Education, 2004(1), 3-31

Grossen, M., & Bachmann, K., (2000). 'Learning to collaborate in a peer-tutoring situation: Who learns? What is learned?', European Journal of Psychology of Education, XV (4), 497-514.

James, M. (2006) Assessment teaching and theories of learning. In J. Gardner (Ed.) Assessment and Learning (pp. 47-60) London: Sage Publications.

Knight, P., & Yorke, M. (2003). Assessment, learning and employability. Buckingham: Open University Press.

Kutti, K. (1996). Activity Theory as a Potential Framework for Human-Computer Interaction. In B. A. Nardi (Ed.), Context and Consciousness: Acitivity Theory and Human-Computer Interaction (pp. 17-44). Cambridge, MA: MIT Press.

Laurillard, D. (1993). Rethinking University Teaching: A Framework for the Effective Use of Educational Technology. London: Routledge

Lave, J., & Wenger, E. (1991). Situated Learning: Legitimate Peripheral Participation. Cambridge, UK: Cambridge University Press.

Littleton, K., Miell, D. & Faulkner, D. (Eds.) (2004) Learning to Collaborate, Collaborating to Learn, New York: Nova Science.
McKillop, C. (2004). 'StoriesAbout... Assessment': supporting reflection in art and design higher education through on-line storytelling Paper presented at the 3rd International Narrative and Interactive Learning Environments Conference (NILE 2004), Edinburgh, Scotland.

Miell, D. & Littleton, K. (Eds.) (2004) Creative collaborations, London: Free-Association Books.

Murphy, P. (2000). 'Understanding the process of negotiation in social interaction', in Joiner. R., Littleton, K., Faulkner, D. & Miell, D. (Eds.) Rethinking collaborative learning, London: Free Association Books.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. Studies in Higher Education, 31(2), 199218.

Norman, D. (1988). The psychology of everyday things. New York: Basic Books.

Pask, G. (1976). Conversation theory: applications in education and epistemology. Amsterdam: Elsevier.

Phye, G.D. and Bender, T. (1989) 'Feedback complexity and practice: response pattern analysis in retention and transfer', Contemporary Educational Psychology, Vol. 14, pp.97–110

Pitcher, N., Goldfinch, J. and Beevers, C. (2002) 'Aspects of computer based assessment in mathematics', Active Learning in Higher Education, Vol. 3, No. 2, pp.19–25.

Rheingold, H (2002) Smart Mobs - The Next Social Revolution, Cambridge, Mass, USA: Perseus.
Rowntree, D. (1977) Assessing Students: How shall we know them? Kogan Page, London

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. Instructional Science, 18, 119-144.

Simpson, O. (2003) Student retention in Online, Open and Distance Learning. Kogan-Page. ISBN 0-7494-3999-8

Thomas, P., Price, B., Paine, C. and Richards, M. (2002) 'Remote electronic examinations: student experiences', British Journal of Educational Technology, Vol. 33, No. 5, pp.537–549.

Vogiazou, Y., Eisenstadt, M., Dzbor, M. and Komzak, J. (2005) Journal of Computer Supported Collaborative Work.
Watt, S., Whitelock, D., Beagrie, C., Craw, I. Holt, J., Sheikh, H. & Rae, J. (2006) Developing open-source feedback tools. Paper presented at ELG Conference, Edinburgh 2006.

Whitelock, D. (1999) 'Investigating the role of task structure and interface support in two virtual learning environments', Int. J. Continuing Engineering Education and Lifelong Learning, Special issue on Microworlds for Education and Learning, Guest Editors Darina Dicheva and Piet A.M. Kommers, Vol. 9, Nos 3/4, pp. 291-301. ISSN 0957-4344.

Whitelock, D., Romano, D., Jelfs, A. and Brna, P. (2000) 'Perfect Presence: What does this mean for the design of virtual learning environments?', in Selwood, I., Mikropoulos, T., and Whitelock, D. (eds) Special Issue of Education & Information Technologies: Virtual Reality in Education, Vol. 5, No. 4, December 2000, pp. 277-289, Kluwer Academic Publishers, ISSN 1360-2357.

Whitelock, D. and Raw, Y. (2003) 'Taking an electronic mathematics examination from home: what the students think', in C.P. Constantinou and Z.C. Zacharia (Eds). Computer Based Learning in Science, New Technologies and their Applications in Education, Vol. 1, Nicosia: Department of Educational Sciences, University of Cyprus, Cyprus, pp.701–713, ISBN 9963-8525-1-3.

Whitelock, D., & Brasher, A. (2006). Developing a roadmap for e-assessment: which way now? Paper presented at the 10th International Computer Assisted Assessment Conference, Loughborough University.

Whitelock, D. & Watts, S. (2007) e-Assessment: How can we support tutors with their marking of electronically submitted assignments? Ad-Lib Journal for Continuing Liberal Adult Education, Issue 32, March 2007. ISSN 1361-6323
Yorke, M. (2003). Formative assessment in higher education: moves towards theory and the enhancement of pedagogic practice. Higher Education, 45(4), 477-501.

The author:

Denise Whitelock
The Open University
Walton Hall
Milton Keynes, MK7 6AA, UK

E-Mail: d.m.whitelock@open.ac.uk
WWW: http://open.ac.uk

Dr Denise Whitelock currently directs the OU's Computer Assisted Formative Assessment project (CAFA) which has three strands: (a) building a suite of tools for collaborative assessment; (b) working with different Faculties to create and evaluate different types of formative assessments; (c) investigating how the introduction of a Virtual Learning Environment (Moodle) is affecting the development of formative assessments in the Open University. All these e-assessment projects demonstrate the synergy between Denise's research and practical application within the Open University.

She has also led the eMentor project, which built and tested a tutor mentoring tool for the marking of tutors' comments on electronically submitted assignments. This project received an OU Teaching Award. Denise is a member of the Educational Dialogue Research Unit (EDRU) Research Group and Joint Information Systems Committee (JISC) Education Experts Group. In November 2007 she was elected to the Governing Council of the Society for Research into Higher Education.

Technology in the service of 21st century learning and assessment

Martin Ripley
Independant Consultant

Abstract:
This article examines the potential role of e-assessment as a catalyst for change in learning and education. The article focuses on recent developments in large-scale e-assessment policy and practice in the UK as well as discussing the ways in which schools can use technology and assessment to support and transform learning in the 21st century. The article suggests that, although there are some encouraging initiatives in e-assessment in the UK, there is not yet a strategic approach at a national level to the further adoption of e-assessment. Lacking this strategic approach, it is hard to see how e-assessment will scale from isolated islands of excellence to a more coherent service. The article contains three sections: (1) The policy framework for e-assessment, which summarises the major policies that relate to e-assessment; (2) aspects and benefits of e-assessment, which provides a description of what counts as e-assessment and the major benefits; (3) Research evidence underpinning e-assessment developments, which summarises major research evidence for the efficacy of e-assessment.

"… Technology can add value to assessment practice in a variety of ways … e-assessment in fact is much more than just an alternative way of doing what we already do." (JISC 2006b, p7)

Introduction

Consider the following two accounts: In 2006 one of the UK's largest awarding bodies, the Assessment and Qualifications Alliance (AQA), completed its first trial of computer-delivered assessment at GCSE. The approach taken by AQA was to create a closed-response computer-delivered test as one component of a science GCSE. The on-screen test was created by selecting suitable materials from past paper-based tests. The AQA pilot was critically reviewed by the national media, who were sceptical of the value of multiple-choice testing. Jonathan Osborne, Professor of Science Education at King's College London said: 'How is this going to assess pupils' ability to express themselves in scientific language, a major aspect of science?' The Times article expressed strong doubt regarding the educational value of this approach to testing, a view shared by many educators in the UK. (The Times Online, 2006)

Over the past decade, there has been unprecedented enthusiasm for the potential of technology to transform learning. The Department for Education and Skills (DfES) has provided significant sums of money for schools to purchase computer equipment and networks, to buy content and management systems. There have been nationwide training and development programmes for teachers and headteachers. And yet, by 2007, most informed commentators have estimated that fewer than 15% of schools in England have embedded technology in their teaching and learning. Ofsted reported that none of the schools in their 2005-06 inspections had embedded ICT. (Ofsted 2006, p78)

The first report reflects a widely held perception that technology 'dumbs down' education and learning. In this view, e-assessment is often perceived to involve multiple-choice testing. The second report reflects a vision of learning in the 21st century (albeit as yet unrealised) which uses technology to personalise learning, with learners increasingly in control of their own learning. In this view, e-assessment is seen as a catalyst for change, bringing transformation of learning, pedagogy and curricula.

Assessment embodies what is valued in education. Assessment – whether in the form of examinations, qualifications, tests, homework, grading policies, reports to parents or what the teacher praises in the classroom – sets the educational outcomes.

To meet the educational challenges of the 21st century assessment must embody the 21st century learning skills such as self-confidence, communication, working together and problem solving. In addition, assessment must support learners' analysis of their own learning and it must support constructivist approaches to learning.

Defining e-assessment

For the purposes of this chapter, a broad definition of e-assessment is needed:

- e-assessment refers to the use of technology to digitise, make more efficient, redesign or transform assessment.
- assessment includes the requirements of examinations, qualifications, national curriculum tests, school based testing, classroom assessment and assessment for learning.
- the focus of e-assessment might be any of the participants within assessment processes – the learners, teachers, school managers, assessment providers, examiners, awarding bodies (based on JISC 2006a, p43).

Overview

This chapter discusses the ways in which schools can use technology and assessment to support and transform learning in the 21st century. It contains three sections:

- The policy framework for e-assessment, which summarises the major policies that relate to e-assessment.
- Aspects and benefits of e-assessment, which provides a description of what counts as e-assessment and the major benefits.
- Research evidence underpinning e-assessment developments, which summarises major research evidence for the efficacy of e-assessment.

The policy framework for e-assessment

In 2005 Ken Boston, the Chief Executive of the Qualifications and Curriculum Authority (QCA) spoke optimistically of a forthcoming transformation of assessment in which technology was presented as a catalyst for change: 'technology for assessment and reporting is the third of three potentially transformative but still incomplete major reforms' (Boston, 2005).

His speech continued by setting out the agenda in order that technology-enabled assessment might fulfil its potential. He described the following three challenges:

- Reforming assessment (ie, placing more emphasis on assessment for learning, in the classroom, and less emphasis on external examinations);
- Improving the robustness of organisations that supply assessments (ie, ensuring that awarding bodies make the change);
- Leading debate regarding standards and comparability with paper-based ancestors of e-assessments (ie, making sure that transformation is not thwarted by media hype about erosion of standards and 'dumbing down').

Whilst acknowledging the risks and difficult choices for suppliers and adopters of e-assessment, Ken Boston's speech concluded with an enthusiastic call for technology to be used to transform assessment and learning: "There is much less risk, and immensely greater gain, in pursing strategies based on transformational onscreen testing; transformational question items and tasks; total learning portfolio management; process-based marking; and life-long learner access to systemic and personal data. There is no political downside in evaluating skills and knowledge not possible with existing pencil and paper tests, nor in establishing a new time series of performance targets against which to report them." (Boston, 2005).

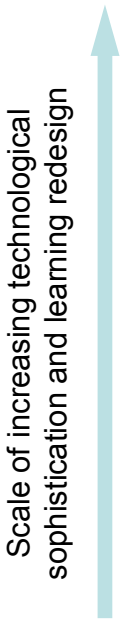
Surprisingly, much of QCA's subsequent policy developments and e-assessment activity has failed to provide the transformation that Ken Boston spoke of. Activity has been regulatory and reactive, not visionary and not providing the necessary leadership. For example, QCA has published two regulatory reviews of issues relating to the use of technology in assessment. The first study focused on issues relating to e-plagiarism (QCA, 2005) and led QCA to establish an advisory team in this area. Some commentators have seen a link between the e-plagiarism study and the subsequent advice from QCA that will lead to significant curtailments in the use of coursework. QCA's second review related to the use of technology to cheat in examination halls (QCA, 2006).

There is a dilemma here for the regulators. At the same time as wanting to demonstrate regulatory control and enhance public confidence in examination systems, the regulatory bodies have wanted to bring about transformation. So, while QCA has been urged to consider banning digital devices, projects (like eSCAPE – see below) have been demonstrating the improvements to assessment that those very same devices can bring.

Aspects and benefits of e-assessment

To understand the contribution of technology through e-assessment we must understand the ways in which it redefines the relationship among learning, the curriculum, pedagogy and assessment. At its most straightforward, e-assessment replicates paper-based approaches to testing. For example, there are several commercially available products which supply national curriculum test materials on screen and on CD-rom, most of which consist of libraries of past test papers. At the other end of the spectrum, however, e-assessment changes pedagogy and assists students in taking responsibility for their learning. It extends significantly our concept of what counts as learning in the classroom, and it supports out of school learning.

Figure One sets out the range of ways in which different types of e-assessment product support different aspects of learning.
In the most sophisticated examples of digital learning, assessment is blended so well with learning that the two become indistinguishable.



Type of product	Examples	Classroom benefits	Issues
Digital learning space	*Digital brain* *UniServity Connected Learning Community (CLC)*	Learner driven Constructivist Encourages learners to review progress in learning	Few applications Most become diverted into mechanical e-portfolios and content rich learning platforms
21st century higher order skills	*KS3 ICT tests* *World Class Tests* *Maths Grid Club?*	Drives 21st century curriculum; higher order skills; communication and problem solving; focus on applying skills and knowledge	Very few resources; often focused on mathematics; require schools to welcome the curriculum changes that this brings
Classroom-based handheld technology	*Wolverhampton's Learn2Go pilot* *Promethean Activpad*	Every pupil has constant access to the technology, enabling true embedding; supports the characteristics of assessment for learning	Emerging technology, not yet established; requires significant energy and commitment from a school to make this work
E-portfolios and related assessment activities	*Measuring Pupil Performance (MPP)* *MAPS, from TAG Learning*	Pupil can 'drive' the assessment; assessment criteria are visible;	
Drill and review	*Jelly James*	Can focus on misconceptions and weaknesses	
Databases of past test papers	*Testbase* *Exampro* *Trumptech* *TestWise*	Provides the teacher with highest quality assessment items, that can be flexibility arranged to create focused tests	Often restricted to test format, but can still offer good diagnostic capability
Analysis tools	*Pupil Tracker*	Aids preparation for tests and examinations, supports the focus on school performance tables	Can be mechanical and results focused, not learning focused

Figure 1: e-assessment products to support learning

An example of strategic development at local authority level can be found in the Wolverhampton system, which has three components. Virtual Workspace is an 'open all hours' learning resource for 16-19 year olds. It provides students with mentors and tutors, able to respond to requests for help by email or telephone. Students have access to on-line course material, and technical training and support are available for school staff. For the second component, Area Prospectus, all 14-16 providers in the area have agreed to enable learners to take courses across a range of institutions. To make this possible they use common names for courses and have designated one day of the week when learners can physically move to other institutions to attend lessons. The third component is a piece of software called My i-plan which records what students are planning to do and how they are progressing. Importantly, it operates through a system of dual logins, providing students with a degree of control and ownership.

Wolverhampton's work is preparing the local authority for the effects of national policy changes that will transform the face of secondary schooling. Those policy changes will provide learners with more flexibility in

where and when they learn, and will require modern assessment systems able to follow the learner and accredit a wide range of evidence of learning. e-portfolios and e-assessment have a fundamental role to play in joining learning with assessment and enabling the learner to monitor progress.

There are a number of compelling reasons why school leaders should consider e-assessment:

- It has a positive effect on motivation and performance

Strong claims are made for the positive effect of technology on pupils' attitudes to learning. E-books have been found to increase boys' willingness to read and improve the quality of their writing (Perry, 2005). Anecdotal evidence suggests that the concentration and performance of even our youngest learners improve when they are using technology. Adult learners self-labelled as school and exam failures have said that e-assessment removes the stress and anxiety they associate with traditional approaches to examinations.
The Learn2Go project in Wolverhampton has experimented with the use of handheld devices in primary and secondary schools (Whyley, 2007). Already this project has demonstrated significant improvements in children's self-assessment, motivation and engagement with the curriculum, including in reading and mathematics. The work is now also claiming evidence that these broad gains translate into improvements in children's scores on more traditional tests.

- It frees up teacher time

Well managed, e-assessment approaches can certainly free up significant amounts of teacher time. Some e-assessment products provide teachers with quality test items and teachers can store and share assessments they create. Where appropriate, auto-marking can enable a teacher to focus on analysis and interpretation of assessment results.

- High quality assessment resources

High quality, valid and reliable assessments are notoriously difficult to design. e-assessment resources provide every teacher with access to high quality materials – whether as a CD-rom containing past questions from national examinations, or as a database of classroom projects with marking material for standardisation purposes, or as websites with interactive problem solving activities.

- Provides rich diagnostic information

E-assessment applications are beginning to provide learners and teachers with detailed reports that describe strengths and weaknesses. For example some examinations provide the student not only with an instant result but also with a report setting out any specific areas for further study; some early reading assessments provide the teacher with weekly reports and highlight children whose progress might be at risk.
There is a distinction to be drawn between the genuinely diagnostic and learner focused reports that some software provides, versus the 'progress tracking' reports available through other products. The purpose of progress tracking reports is to ensure that pupils achieve targeted national curriculum and GCSE results, and they are quite different from diagnostic reporting.

- Flexible and easy to use

One of the strongest arguments in favour of e-assessment is that it makes the timing of assessment flexible. Formal assessments can be conducted when the learner is ready, without having to wait for the annual set day. Many providers of high-stakes assessments nowadays require no more than 24 hours notice of a learner wanting to sit an examination. Diagnostic assessments can be provided quickly, and at the relevant time.

- Links learning and assessment, empowering the learner

One of the core principles of assessment for learning is that assessment should inform learning. The learner is therefore the prime intended audience for assessment information. E-assessment tools can provide ways of achieving this - for example, e-portfolios should always enable the learner to collect assessment information, reflect on that information, and make decisions (with the support of a teacher when appropriate) about future learning steps.

- Assessment of high order thinking skills in ways not possible with paper-and-pencil testing

World Class Tests, developed by QCA, are designed to assess higher order thinking skills in mathematics and problem solving for students aged 9-14. They are one of best examples of computer-enabled assessments and have set expectations for the design of on-screen assessment.

- It is inevitable

An increasing range of assessments is being developed for use on computer. Few of us can apply for a job without being required to complete an on-screen assessment; many professional examinations are now administered on-screen; whole categories of qualification (such as key skills tests) are now predominantly administered on-screen. Awarding bodies are already introducing e-assessments into GCSEs and A-levels. The QCA has set a 'Vision and Blueprint', launched in 2004 by the then Secretary of State for Education and Skills, Charles Clarke which heralds significant use of e-assessment by 2009 (QCA 2004). The question is not so much whether a school should plan for e-assessment, but why a school would wish to wait and delay.

These descriptions of the benefits to teachers and learners of e-assessment are compelling. It is also clear that the primary benefit of e-assessment is that it supports effective classroom learning in accordance with the characteristics of assessment for learning.

Developments and Research

There have been few studies of technology's impact on learning of technology. Some studies have found that frequent use of technology in school and at home correlates with improved examination performance on the traditional, paper-based tests used at key stage 3 and GCSE (Harrison et al., 2002). However, it is not clear whether it is technology that makes the difference, or whether technology tends to exist in families and social groups more likely to do well in traditional measures of educational performance. This research should be compared with the Education Testing Service study referred to below, which found different effects for some students taking tests on computer.

Developments and research in e-assessment are in their early days, but a growing body of evidence is accumulating, some of which is reviewed below.

- Scotland

The e-assessment work of The Scottish Qualifications Authority (SQA) includes Pass-IT (which investigated how e-assessments might enhance flexibility, improve attainment and support teaching and learning) and guidelines on e-assessment for schools (Scottish Qualifications Authority, 2005). E-assessment has been used in high stakes external examinations including Higher Mathematics and Biotechnology Intermediate 2. The Scottish OnLine Assessment Resources project is developing summative online assessments for a range of units within Higher National qualifications. SQA is developing three linked on-screen assessment tools for Communication, Numeracy and IT, and is also investigating the use of wikis and blogs for assessment (Scottish Qualifications Authority, 2006). The SCHOLAR programme developed within Heriot-Watt University provides students with an on-line virtual college designed to help students as they progress between school, college and university.

- eSCAPE

The eSCAPE project led by Richard Kimbell at the Technology Education Research Unit (TERU) at Goldsmiths College, and by TAG Learning, has focused on GCSE design and technology. Its purpose has been to design short classroom-administered assessments of students' ability to create, prototype, evaluate and communicate a solution to a design challenge. In the eSACPE project:

- students work individually, but within a group context, to build their
- design solution;
- each student has a pda, with functionality enabling them to video, photograph, write documents, sketch ideas and record voice messages;
- at specified points in the assessment, students exchange ideas and respond to the ideas of others in the group;
- at the end of the assessment, students' portfolios are loaded to a secure website, through which human markers score the work.

A report of phase 2 (TERU, 2006) described the 2006 pilot in which over 250 students completed multi-media e-portfolios and submitted these to TERU, who had trained a team of markers to mark the e-portfolios on screen. The assessment efficacy and the robustness of the technology have proven highly satisfactory. Students work well with the technology and rate the validity of the assessment process positively.

The eSCAPE assessment uses an approach to marking known as Thurstone's graded pairs. Human markers rank order students' work, working through a series of paired portfolios. For each pairing, the markers record which of the two pieces of work is better. Based on the positive evaluation findings of this approach, QCA has encouraged further development and Edexcel is planning to apply the eSCAPE approach.

- Key stage 3 ICT tests

The key stage 3 ICT test project is one of the largest and most expensive e-assessment developments in the world. The original vision for the tests involved the creation of an entire virtual world, with students responding to sophisticated problems within the virtual world. For this vision to work, the project needed to deliver successful innovation on several fronts, for example:

- developing the virtual world;
- developing a successful test form within the virtual world, including a test that could reliably measure students' use of their ICT skills;
- developing a new psychometric model;
- training all secondary schools in the technical and educational adoption of the tests;
- redesigning the teaching of ICT.

However, the full range of planned innovation has not been delivered. In particular, the tests have adopted more traditional approaches to test design, and teachers generally have not been persuaded that the tests reflect improved practice in ICT teaching. Nevertheless, the project is one of the most evaluated e-assessment projects (see for example QCA, 2007b) and is an excellent source of information for other organisations considering the development of innovative forms of e-assessment. QCA is now however seeking to develop the underlying test delivery system into a national infrastructure, making this available to test providers for the purposes of delivering large-volume, high stakes tests to schools. This is known as Project Highway.

- 21st Century Skills Assessments

Margaret Honey led a world-wide investigation into the existence and quality of assessments in key areas of 21st century learning (Partnership for 21st Century Skills, 2006). Her report highlighted 'promising assessments' including England's key stage 3 ICT tests. She found that although educators in many countries agree that ICT skills are core learning skills, it is only in the UK that substantial progress has been made in developing school-based assessment of ICT. Honey was unable to find any measures that address students' understanding of global and international issues, although she reported that in the US assessment of civic engagement is quite well established. Also in the US, an assessment of financial, economic and business literacy is currently being developed and will become mandatory for students in year 12.

- Technical measurement issues

In 2005 the Education Testing Service (ETS) published the findings of a large scale comparison of paper-based and computer-delivered assessments (National Centre for Education Statistics, 2005). The empirical data were collected in 2001 and involved over 2,500 year 8 students who completed either mathematics or writing assessments. The traditional, paper-based tests were migrated to screen format, with little or no amendment made for the purpose of screen delivery. In mathematics the study found no significant differences between performance on paper and on screen, except for those students reporting at least one parent with a degree. These students performed better on paper. The study also found no significant differences in writing, expect for students from urban fringe/large town locations. Again these students performed better on paper than on screen. The purpose of the ETS study was to investigate the efficacy of migrating existing eight grade tests from paper to screen. The study concluded that this was achievable, although with some loss of unsuitable test items. The study did not examine the issue of whether such a migration would be educationally desirable nor whether computer-delivered tests should include an aim to transform test content.

- Market surveys

Thompson Prometric commissioned two reviews of e-assessment issues involving all UK awarding bodies (Thompson Prometric 2005 and 2006). They achieved high levels of participation and revealed that the majority of awarding bodies are actively pursuing e-assessment, although often without senior executive or strategic involvement. The studies made clear the remarkable agreement between awarding bodies regarding the

benefits of e-assessment, which included learner choice, flexibility and on-demand assessment. There was also agreement between most awarding bodies regarding the major issues – authentication, security, cost and technical reliability.

Conclusions

In the words of Ken Boston, there is much to be gained by considering the transformative potential of e-assessment. This chapter has sought to describe the importance of linking e-assessment to strategic planning for the future of learning, as well as identifying a number of ways in which e-assessment can support effective learning in the classroom.

In a significant recent development, an eAssessment Association (eAA) has been created by Cliff Beevers, Emeritus Professor of Mathematics at Heriot-Watt University. (See eAssessment Association.) The group was launched in March 2007, with involvement from industry, users and practitioners. The eAA aims to provide members with professional support, provide a vision and national leadership of e-assessment, and publish a statement of good practice for commercial vendors. It is to be hoped that the eAA will play a significant role in encouraging the assessment community to make use of technology to improve assessment for learners.

Further reading

MacFarlane, A. (2005) Assessment for the digital age. Available from: http://www.qca.org.uk/ibraryAssets/media/11479_mcfarlane_assessment_for_the_digital_age.pdf Accessed 10th July 2007

Microsoft (2005), Wolverhampton City Council Mobilises Learning to Give Students Access to Anywhere, Anytime Education. Available from http://download.microsoft.com/documents/customerevidence/8097_Wolverhampton_Final.doc Accessed 10th July 2007

Naismith, L., Lonsdale, P., Vavoula, G. and Sharples, M. (2004) Literature Review in Mobile Technologies and Learning. Futurelab Report Series no. 11.

Pass-IT (2007) Pass-IT - Project on Assessment in Scotland using Information Technology. Information available from www.pass-it.org.uk

Ridgway, J. (2004) e-Assessment Literature Review. Available from http://www.futurelab.org.uk/resources/publications_reports_articles/literature_reviews/Literature_Review204 Accessed 10th July 2007

Ripley, M. (in print) Futurelab e-assessment literature review – an update. To be available from www.futurelab.org.uk

Scholar Programme available at http://scholar.hw.ac.uk/

Scottish Qualifications Authority (2005) SQA Guidelines on e-assessment for Schools, SQA Dalkeith, Publication code: BD2625, June 2005, www.sqa.org.uk/files_ccc/SQA_Guidelines_on_e-assessment_Schools_June05.pdf Accessed 10th July 2007

References

Boston, K. (2005) Strategy, Technology and Assessment. Speech delivered to the Tenth Ann ual Round Table Conference, Melbourne, October 2005. Available from http://www.qca.org.uk/qca_8581.aspx Accessed 10th July 2007

DfES (2005) Harnessing Technology. Available at http://www.dfes.gov.uk/publications/e-strategy/ Accessed 10th July 2007

DfES (2006) 2020 Vision: Report of the Teaching and Learning in 2020 Review Group. London: DfES. Available at http://www.teachernet.gov.uk/educationoverview/briefing/strategyarchive/whitepaper2005/teachingandlearning2020/ Accessed 10th July 2007

e-Assessment Association, www.e-assessmentassociation.com Accessed 10th July 2007

Harrison, C., Comber, C., Fisher, T., Haw, K., Lewin, C., Lunzer, E., McFarlane, A., Mavers, D., Scrimshaw, P., Somekh, B. and Watling, R. (2002) ImpaCT2: The Impact of Information and Communication Technologies on Pupil Learning and Attainment. Available at: www.becta.org.uk/research/impact2 Accessed 10th July 2007

JISC (2005) Innovative Practice with e-Learning: A good practice guide to embedding mobile and wireless technologies into everyday practice. Available from the JISC website: www.jisc.ac.uk/uploaded_documents/InnovativePE.pdf Accessed 10th July 2007

JISC (2006a) e-Assessment Glossary – Extended. Available at http://www.jisc.ac.uk/uploaded_documents/eAssess-Glossary-Extended-v1-01.pdf Accessed 10th July 2007

JISC (2006b) Effective Practice with e-Assessment: An overview of technologies, policies and practice in further and higher education. Available from: http://www.jisc.ac.uk/media/documents/themes/elearning/effprac_eassess.pdf Accessed 10th July 2007

National Centre for Education Statistics (2005) Online Assessment in Mathematics and Writing. Reports from the NAEP Technology-Based Assessment Project, Research and Development Series. Available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2005457 Accessed 10th July 2007

O'Brien, T.C. (2006) Observing Children's Mathematical Problem Solving with 21st Century Technology. Available from http://www.handheldlearning.co.uk/content/view/24/2/ Accessed 10th July 2007

Ofsted (2006) The Annual Report of Her Majesty's Chief Inspector of Schools, 2005-06. Available from http://www.ofsted.gov.uk/assets/Internet_Content/Shared_Content/Files/annualreport0506.pdf Accessed 10th July 2007

Partnership for 21st Century Skills (2006), The Assessment Landscape, available for http://www.21stcenturyskills.org/images/stories/otherdocs/Assessment_Landscape.pdf Accessed 10th July 2007

Perry, D. (2005) Wolverhampton LEA 'Learn2Go' Mobile Learning PDAs in Schools Project, Evaluation Phase 1, End of First Year Report. Available from http://www.learning2go.org/pages/evaluation-and-impact.php Accessed 10th July 2007

QCA (2004) QCA's e-assessment vision. Available from http://www.qca.org.uk/qca_5414.aspx Accessed 10th July 2007

QCA (2005) A review of GCE and GCSE coursework arrangements. Available from http://www.qca.org.uk/qca_10097.aspx Accessed 10th July 2007

Underwood, Jean (2006) Digital Technologies and dishonesty in examinations and tests. Published by QCA, 2005. Available from http://www.qca.org.uk/qca_10079.aspx Accessed 10th July 2007

QCA (2007a) Regulatory Principles for e-assessment. Available from http://www.qca.org.uk/qca_10475.aspx Accessed 10th July 2007

QCA (2007b) A review of the Key Stage 3 ICT test at http://www.naa.org.uk/naaks3/361.asp Accessed 10th July 2007

Thompson Prometric (2005) Drivers and Barriers to the Adoption of e-Assessment for UK Awarding Bodies. Thompson Prometric, 2005

Thompson Prometric (2006) Acceptance and Usage of e-Assessment for UK Awarding Bodies. Available at http://www.prometric.com/NR/rdonlyres/eegssex2sh6ws72b7qjmfd22n5neltew3fpxeuwl2kbvwbnkw2ww2jbodaskpspvvqbhmucch3gnlgo3t7d5xpm57mg/060220PrintReady.pdf Accessed 10th July 2007

TERU (2006) The phase 2 eSCAPE report will be available on the TERU site shortly: http://www.teru.org.uk Accessed 10th July 2007

The Timesonline (2006), Select one from four for a science GCSE by Tony Halpin. http://www.timesonline.co.uk/article/0,,2-2219509,00.html Accessed 10th July 2007

Whyley, D. (2007) Learning2Go Project. Information available from http://www.learning2go.org/ Accessed 10th July 2007

The author:

Martin Ripley
3 Hampstead West
224 Iverson Road
West Hampstead
London NW6 2HX
E-Mail: martin.ripley1@btinternet.com

Martin Ripley is a leading international adviser on 21st century education and technology. He is co-founder of the 21st Century Learning Alliance, and owner of World Class Arena Limited. He is currently working with a number of public sector organisations and private sector companies in Asia, Europe and the USA.

In 2000 Martin was selected to head the eStrategy Unit at England's Qualifications and Curriculum Authority (QCA). He won widespread support for his national Vision and Blueprint for e-assessment. He led the development of one of the world's most innovative tests – a test of ICT for 14 year-old students.

Martin has spent 15 years in test development. He has been at the heart of innovation in the design of tests in England: he developed England's national assessment record for 5 year-old children; he developed England's compulsory testing in mathematics and science for 11 year-olds; he introduced the UK's first national, on-demand testing programme; to critical acclaim, he developed World Class Tests - on-screen problem solving tests that are now used world-wide as a screening tool for gifted students. These problem solving tests are now sold commercially around the world, including in China and Hong Kong.

Stimulating innovative item use in assessment

René Meijer
University of Derby

Abstract:
Assessment is one of the most powerful influences on learning. Key characteristics that have been identified as contributing significantly to learning are the use of feedback, task authenticity, and the adaptation of teaching based on assessment outcomes. The use of technology has the potential to dramatically enhance our ability to implement these characteristics. This paper is based on a first exploration of the requirements for computer aided assessment in preparation of the replacement of the assessment system currently in use at the University of Derby. It describes a conceptual model for assessments and assessment systems based on the pedagogical requirements of the university. The key characteristic of this model is that it moves away from the question and the test as the defining concepts around which standards, solutions and, as a result, our thinking has become structured. It is a model that is closer to, and as a result more integrated in, learning design. The value of this conceptual model is threefold:

- *It encourages a pedagogical perspective on developing computer aided assessment*
- *It removes barriers to entry for new and innovative assessment items and practice to be developed and shared*
- *It provides a more flexible system architecture for supporting a wide range of assessment practices.*

The value of assessment

Assessment can be one of the most powerful influences on learning. The expectations of, or perceptions on, what will be assessed are amongst the primary motivators of teaching and learning. It is therefore important that assessment tasks are authentic (William, 1994), and are not just a distant proxy for, or subset of the outcomes desired. This is not just important in the context of the summative assessment, but perhaps even more so in the formative domain. Here the assessment, aside from providing an opportunity to reflect on progress and attainment, should also engage the student with the appropriate amount and type of learning (Gibbs 2004). This learning is supported best by timely and specific feedback (Hattie,1987). Ideally assessments do more then provide information and guidance. They become continuous; integrated into learning and adapting the curriculum to match the developing needs of the learner (Black, 2001).

Providing continuous specific feedback, personalised learning and rich authentic tasks for learners is often constrained by the amount of time and preparation these require. When we have the resources for individual tutorials these principles can find their way into teaching relatively easily. In the more common setting of a less advantageous tutor to learner ratio however, doing justice to these ideals becomes more difficult. Technology, when used appropriately, can dramatically increase our ability to implement these principles.

The value of assessment technology

Randy Bennett describes 3 generations of computer-based assessment (Bennett, 1998), broadly mapping onto the normal adoption phases for technology (substitution, innovation and transformation):

- The 1st Generation automates the existing process without reconceptualising it (e.g. multiple choice examinations). Assessment technology in the first generation of the model can enhance the ability to give timely feedback. By automation of the predictable it can limit the need to personally engage with large volumes of repetitive feedback, allowing the teacher to focus on more specific and complicated guidance.
- The 2nd Generation uses multimedia technology to assess skills in ways that were not previously possible (e.g. simulations). This will allow us to explore and use new modes of testing that will dramatically increase the authenticity of the assessments
- With generation 'R' assessment will become indivisible from instruction, with high stakes decisions being made on many assessments. Generation 'R' assessment obviously links directly to the integration into, and subsequent adaptation of, learning and teaching.

At the time of Bennett's paper most assessment systems could be classified as belonging to the early phases of generation 1. They primarily seemed to satisfy our increasing desire to generate large numbers of quantitative measurements with the highest possible efficiency. For the University of Derby, which is characterised by a very broad and diverse curriculum with relatively small cohort sizes, efficiency and scalability were never a major part of the business case. The value of using technology in assessment was mainly pedagogical. Initially this added value was sought in the use of media and simulations, requiring at least a second generation implementation. This requirement was one of the important factors that lead to the development of the Tripartite Interactive Assessment Delivery system (TRIADS) as one of the outcomes of a project funded by the Higher Education Funding Council for England (HEFCE) and delivered by the University of Liverpool, the University of Derby and The Open University. The system has been very successful, and is still used as the primary assessment system by the Centre for Interactive Assessment Development (CIAD) at the University of Derby.

Assessment practice is developing rapidly. There is an increasing demand for the support of more integrated and continuous assessment, and collaborative work and peer review processes. Additionally, while TRIADS is still pedagogically adequate, its code base is about ten years old and needs a refresh, in particular now that Adobe has announced to stop supporting the underlying technology: Authorware 7. It is against this background that the University of Derby is investigating the options for replacing TRIADS. Unfortunately it seems that 10 years after Bennett's paper computer based assessment is still stuck in the very early phases of his first generation. Systems and standards are still based around a monolithic question-based model of assessment. When discussing assessment, and in particular computer aided assessment, it is often implicitly assumed that assessments are tests, separate and distinct entities with their own dynamics and structure. This structure in turn mainly revolves around the sequencing of questions. Assessments are thought of as separate events, supported by separate tools, requiring separate standards.

The problem with items

Scalise (2006) describes an assessment item as any interaction with a respondent from which data is collected with the intent of making an inference about the respondent. Following this definition, the defining features of the item are therefore related to process and purpose, not content or structure. Especially if we want to do justice to the principle of authentic assessment, the nature of items should not be restricted. In the domain of learning content we seem to have intuitively grasped this requirement. Standards around learning content, such as SCORM, don't define a structure for the content. It merely describes a standard way of classifying it (through metadata) and it provides an interface to other elements in the learning environment (through the SCORM Application Programming Interface or API). It has standardised how content connects and relates to other elements, not what it consists of.

In the domain of assessment unfortunately different choices were made. The most prominent standard is IMS QTI (Question and Test Interoperability). It is widely (although seldom completely) supported, and its structure is representative of the majority of assessment systems. IMS QTI did focus on item content and structure, resulting in an undesirable but inevitable limitation on the types of items that could be defined within the standard. It did make the standard relatively easy to adopt (although few major suppliers actually managed to do so completely and unambiguously). This pedagogical limitation was one of the primary reasons not to adopt the standard in the TRIADS system, as this would mean taking significant steps backwards in terms of its functionality.

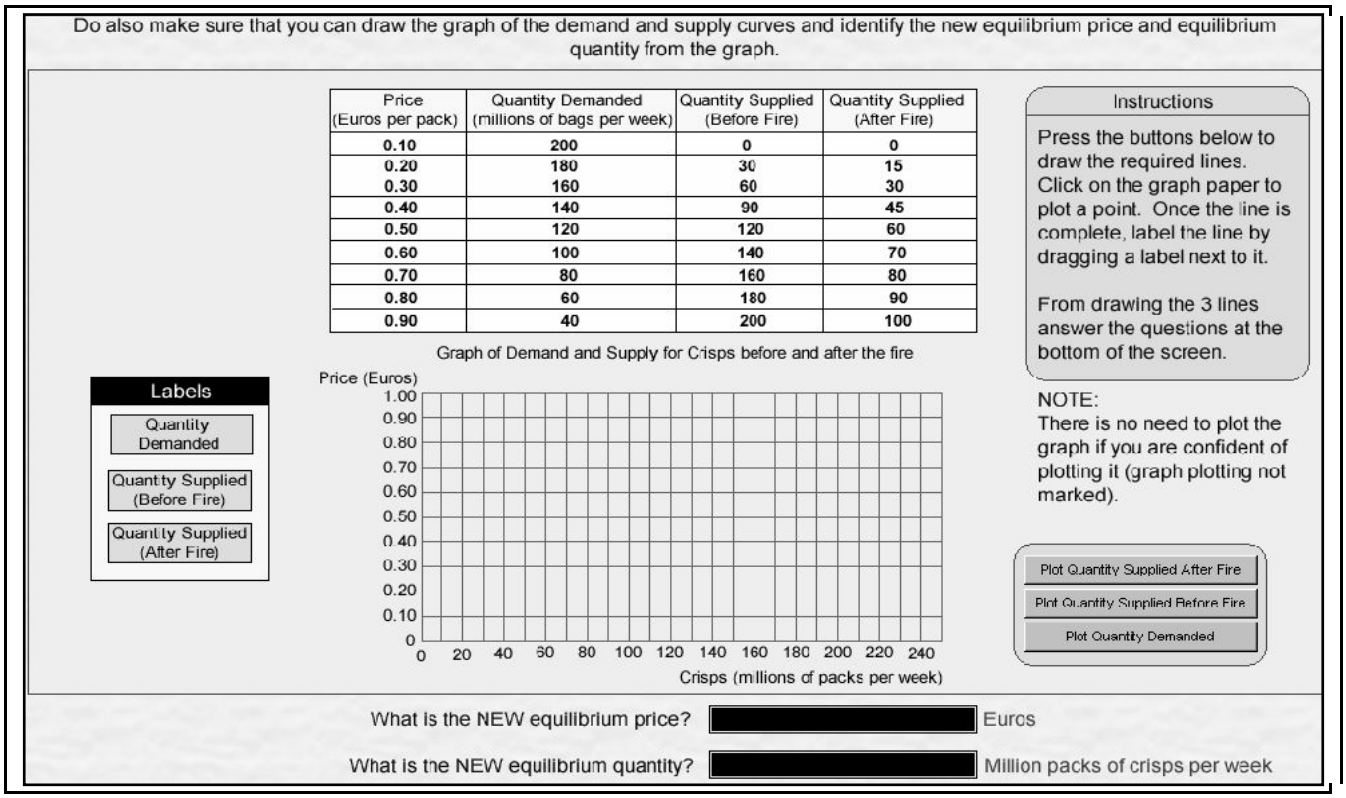


Illustration 1: An example of a complex question type built in TRIADS. Complex types like this were not possible in IMS QTI 1.x. In theory these interactions are supported in IMS QTI 2.x, but in practice these features of the standard will probably not be implemented by any mainstream suppliers.

In the current drafts of version 2.x these limitations have been drastically reduced. However, as the standard is still based on an abstraction of content and structure, this has resulted in a dramatic increase in the complexity of the standard as it tries to cater for more and more variations in item design. No system that I am aware of has adopted the 2.x specification in full, and given its complexity and the limited scope for the uptake of many of the advanced features it seems unlikely for this to ever happen. In particular now that all major suppliers seem to be gathering behind the IMS Common Cartridge, chances are that their implementation of IMS QTI will be limited to the modified version of version 1.2.1 of the standard that is part of the Common Cartridge specification. And so it seems we are stuck within this system in which pedagogical affordances and technological complexity are forever at odds with one another.

Items as widgets

With the advancement of the web as a mature and pervasive platform we are less and less dependent on the development of specific applications to provide us with a runtime environment. Exchange and reuse of content via open or proprietary content standards is being complemented by the exchange of functionality through scripting, Service Oriented Architecture (SOA) or open Application Programming Interfaces (API's). One of the notable developments on the web (although this development is now finding its way back to the desktop, in particular in the domain of open source applications), has been the increasing use of rich and versatile platforms who's functionality is customisable and limited only by the contributions made to it by the community. These contributions are called extensions, plug-ins, gadgets or widgets. Despite the variety in nomenclature, they represent the same basic underlying principle. A widget is a self contained application that is used within a larger application (website, social network or even on the desktop). Examples of these frameworks include Netvibes (http://www.netvibes.com/), Facebook (http://www.facebook.com/) and Google Desktop (http://desktop.google.com/). Widgets interact with the framework in which they run, but also often interact with other information such as weather reports, or stock market information. A similar development can be seen on the desktop, in particular with open source applications such as Firefox (http://www.mozilla.com/en-US/firefox) and games, for instance the use of the scripting language Lua with which the interface of the massively multiplayer online role-playing game 'World of Warcraft' can be extended.

The benefits of adopting a widget or plug-in architecture are threefold:

- Widgets do not need to be limited in their functionality, all that is predefined is how they interact with the framework
- Widget functionality can be relatively self contained, and so widgets could operate outside of a framework (i.e. within learning activities and materials) as well as inside of the framework (i.e. within a separate and distinct assessment)
- Widgets are interchangeable, as long as the framework has implemented a compatible interface (API).

Adopting a widget-like architecture for items would address some of the shortcomings identified earlier. It would allow the degrees of freedom for item design that we desire. Widgets could be as simple as a multiple choice item, or as complex as a 3D simulation. The only thing that needs to be defined is a standardised API to handle the desired exchange of information on the candidates and their responses. When designed right, these widgets could operate either within a regular assessment, but also as stand-alone activities that are integrated within learning materials.
This architecture would also dramatically improve the development and uptake of innovative item types. Most traditional business models focus on the delivery of large volumes of popular solutions, ignoring specialist solutions as the cost of their development and distribution is too expensive compared to the expected revenues. In the current market, where a new item type requires modifications in standards and monolithic software, this effect is strong, which is why most assessment systems are severely limited in the item types they support. In the 'widget-model' however we are no longer restricted by the framework, or the standards, for developing new item types. As a result we can now benefit from what Anderson (2006) has called 'the long tail'. This is the phenomenon whereby low production, stocking and distribution costs allow for a viable market to develop for small-volume solutions.

The problem with tests

Aside from the unsatisfactory limitation on item types, there are other undesirable features of prevalent systems and standards. One of these features is that, while an assessment is clearly an activity, the models defining it seem to be more aligned with object models. Important elements of activities such as roles are missing completely, and so it is impossible to support assessment processes involving multiple stakeholders in collaborative projects, or asynchronous processes such as peer assessment.

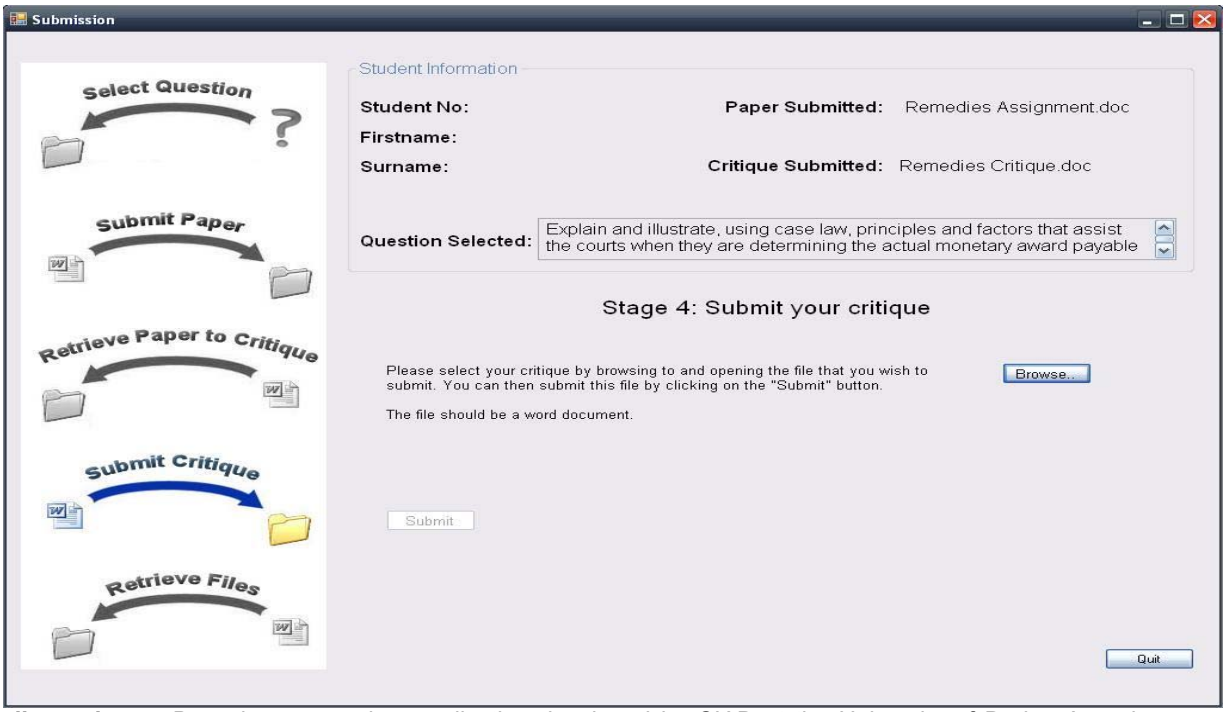


Illustration 2: Bespoke peer review application developed by CIAD at the University of Derby. Asynchronous assessment processes are not well supported by existing systems and standards.

A different model for assessment

A more suitable model for assessments to use as a basis for our solution is that suggested by Allmond (2002). It recognises 4 principle processes that define every assessment. The sequencing and prominence of each of these processes can vary widely depending on the type of assessment that is delivered.

While the model still seems to lean a bit towards the monolithic assessment, it gives the degrees of freedom that we require. An important feature of the model is that it does not assume that al of the components are automated, or even supported by technology. These choices can be made based on the type of solution required. Fully automated multiple choice tests fit the model, but so do those supported by the optical mark reader (where the presentation process is implemented on paper) or an essay task that is submitted via the VLE, but marked and graded by the lecturer.

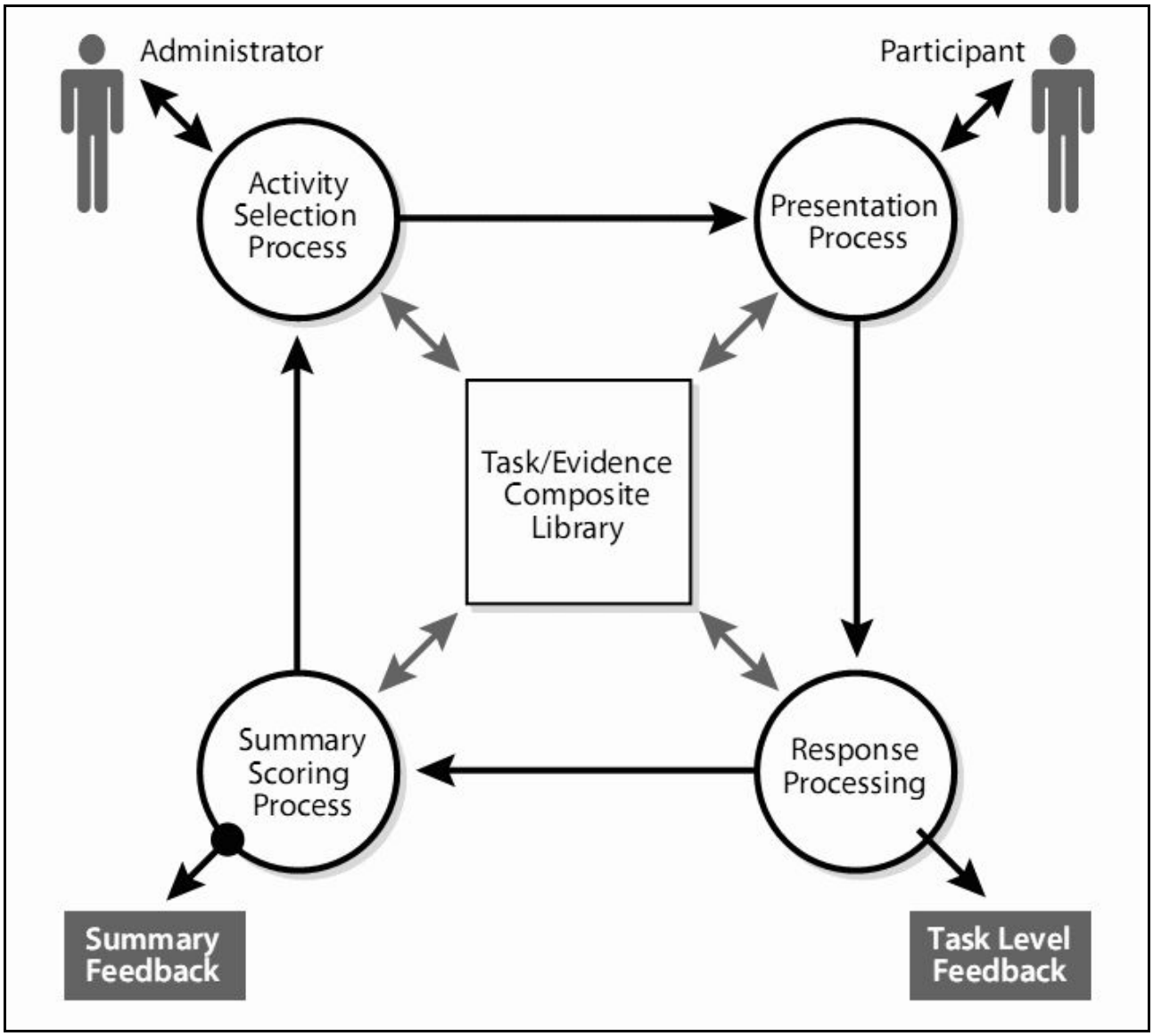


Illustration 3: The four principle processes in the assessment cycle

Composite library

The composite library holds item-widgets, packaged together with the metadata required for selection and use of the items. Existing packaging standards, such as SCORM, could be used to this end, although suggestions by Aroyo (2003) and Plichard (2004) that using semantic web standards (or integrate those standards into SCORM) could possible provide a more flexible and generic solution in the long run deserves serious attention as well.

Activity selection
The activity selection process basically represents the sequencing of the activities. These could be assessment activities from our library, but can also include other learning activities or administrative tasks. In essence this is a process akin to Learning Design, and implementation is probably best done using the relevant standards such as IMS LD. The big advantage of using IMS LD is that it would allow for a truly integrated design of assessment within learning, as explored by Miao (2007).

Presentation
The Presentation Process is the process that presents the task to the participant. This process should largely be handled by the item-widget itself, and its standards are defined by characteristics of the client software that will access the content (for instance a web browser with a flash plug-in). Allowing the item to present itself will avoid having to predefine functionality within delivery systems, which inevitably will lead us back to the current restrictive approach to item design.

Response processing
Response Processing takes the work products from the candidates' response, and records them as observations on the candidates' performance. The basis for the response processing is handled by the item-widget, although a standard will have to be defined for the transferral of the observations to the framework in which the item runs. Elements from IMS QTI (namely around Response Processing) might be reusable in this context.

It is important to separate this process from the item delivery for 3 reasons:

- There are assessment scenarios whereby only 1 of these 2 processes is automated. An essay assignment could be delivered electronically, but marked by hand. Alternatively a questionnaire could be delivered via paper, scanned in with an optical mark reader and marked by computer.

- Access to the responses, in stead of just scores, is crucial for purposes of moderation. If marking schemes need to be adjusted after delivery, responses can be processed again using the new scoring algorithm.

- Access to responses will better support the evaluation of assessment. Understanding exactly which incorrect responses were given can feed back into teaching, or perhaps in modifications of the assessment.

Summary scoring
The Summary Scoring process uses the observations from the response processing to build an aggregate conclusion on the participant's performance. This process is handled by the framework. Standards that could be used to handle this communication are elements of QTI such as the 'OutcomeDeclaration' and 'Outcome Processing'.

Conclusion

The conclusion emerging from this preliminary analysis is that computer-aided assessment is best implemented by extending the tools, systems and standards that we use for learning. With a few enhancements the tools and systems we use to design and deliver learning could also be used to design and deliver assessment. Pedagogically this integration could be crucial in realising the full potential of formative assessment. It would also lower the cost of computer aided assessment, as investments that need to be made in software development, but also the training of staff, can be reduced. If this integrated architecture is combined with the flexibility of the item definition described in this paper, we might finally be on the path to Generation 'R'.

Computer-aided assessment has a tremendous potential to add value to learning and teaching. In order to realise this value it is important that we create systems and standards that provide support for in stead of restrain the degrees of pedagogical freedom that are required. This article has presented some perceived shortcomings in current systems and standards, and suggested an alternative approach that would allow for a viable and flexible implementation of computer aided assessment using existing and mainstream technologies and standards. This approach would significantly lower the barrier for the development and uptake of innovative item types, while at the same time realising a better integration with other learning activities.

References

Advanced Distributed Learning - SCORM®" , http://www.adlnet.gov/scorm/

Anderson, C. (2006) The Long Tail: Why the Future of Business is Selling Less of More, Hyperion.

Allmond, R.G., Steinberg, L.S. & Mislevy, R.J. (2002) Enhancing the Design and Delivery of Assessment Systems: A Four Process Architecture. The Journal of Technology, Learning, and Assessment, 1(5).

Aroyo, L., Pokraev, S. & Brussee, R. (2003) Preparing SCORM for the Semantic Web. In On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. Lecture Notes in Computer Science. Catania Sicily, Italy: Springer Berlin / Heidelberg, p. 621-628.

Black, P. & Wiliam, D. (2001) Inside the Black Box, Raising Standards Through Classroom Assessment. BERA.

Bennet, R.E. (1998) Reinventing Assessment: Speculations on the Future of Large-Scale Educational Testing. Available at: http://www.ets.org/Media/Research/pdf/PICREINVENT.pdf

Gibbs, G. & Simpson, C. (2004) Conditions Under Which Assessment Supports Students' Learning. Learning and Teaching in Higher Education, (1).

Hattie, J.A. (1987) Identifying the salient facets of a model of student learning: a synthesis of meta-analyses, International Journal of Educational Research,vol. 11, pp. 187-212.

IMS Global Learning Consortium: Common Cartridge," http://www.imsglobal.org/commoncartridge.html

IMS Global Learning Consortium: Learning Design Specification," http://www.imsglobal.org/learningdesign/

IMS Global Learning Consortium: IMS Question & Test Interoperability Specification." http://www.imsglobal.org/question/index.html

Miao, Y. et al. (2007) The Complementary Roles of IMS LD and IMS QTI in Supporting Effective Web-based Formative Assessment.

Plichard, P. et al. (2004) TAO, A Collective Distributed Computer-Based Assessment Framework Built on Semantic Web Standards. In In Proceedings of the International Conference on Advances in Intelligent Systems – Theory and Application AISTA2004. Luxembourg. https://www.tao.lu/downloads/publications/AISTA04-paper244-TAO.pdf

Scalise, K. & Gifford, B. (2006) Computer-Based Assessment in E-Learning: A Framework for Constructing "Intermediate Constraint" Questions and Tasks for Technology Platforms. The Journal of Technology, Learning, and Assessment, 4(6).

TRIADS" http://www.triadsinteractive.com/Assessment/triadsdemo/

Wiliam, D. (1994) Assessing authentic tasks: Alternatives to mark-schemes. Nordic Studies in Mathematics Education, 2(1), p.48-68.

The author:

René Meijer
University of Derby
Kedleston Road
Derby
DE22 1GB
United Kingdom

E-Mail: r.meijer@derby.ac.uk
WWW:
http://renesassessment.blogspot.com/

René Meijer is head of the Centre for Interactive Assessment Development (CIAD). His main interests are technology enhanced learning, in particular innovative methods of assessment.

AVAILABLE GUIDELINES AND STANDARDS FOR PSYCHOMETRIC TESTS AND TEST USERS

Prof Dave Bartram
SHL Group plc

Abstract
The paper describes why we need guidelines and standards on tests and test use and why, in particular, we need international agreement on what these should be. The work of the International Test Commission (ITC) is described and the ITC's International Guidelines are reviewed. Various other important national initiatives in Britain, Germany, the Netherlands, Sweden and the USA are described together with the work of the European Federation of Psychologists Associations Standing Committee on Tests and Testing. While there is considerable agreement on what constitutes good practice in test use, there is wide diversity in the ways in which different countries have attempted to implement good practice or regulate test use. The need for guidelines on good practice and for standards for tests and test user competence is ever more urgent in an increasingly global and distributed assessment environment and with the growth in use of computer-based assessment[1].

The issue of setting effective quality standards for the use of psychological assessment is one that has taxed practitioners for many years. Many different approaches have been adopted to try to ensure that tests are used well and fairly and that their results are not misused. The paper will review the case for developing international guidelines on test use, rather than just local ones. It will review some key outcomes of a recent International Test Commission (ITC) survey and describe a number of important ITC Guidelines. Other national and regional developments will be reviewed, including the work of the EFPA Standing Committee on Tests and Testing, the development within Germany of a national process standard for job recruitment and selection (DIN 33430), progress in Britain, Sweden and the Netherlands on the establishment of test institutes for best practice and work in Britain, Norway, Sweden, Finland and the USA on test user qualification.

Why do we need International Guidelines?

In the relatively recent past, it was possible to think of individual countries as 'closed systems'. Changes could be made in terms of best practice, procedures and laws affecting the use of tests in one country without there being any real impact on practice or law in other countries. People tended to confine their practice of assessment to one country and test suppliers tended to operate within local markets - adapting prices and supply conditions to the local professional and commercial environment. This has changed. Individual countries are no longer 'closed' systems. Psychological testing is an international business. Many test publishers are international organisations, selling their tests in a large number of countries. Many of the organisations using tests for selection and assessment at work are multinationals. In each country test suppliers and test users are likely to find differences in practices related to access, user qualification, and legal constraints on test use. Not only are organisations becoming more global in their outlook, so too are individuals.

The increased mobility of people and the opening of access to information provided by the Internet have radically changed the nature of the environment in which we operate. For example, I can now register as a test user with a web-based distributor in the USA and buy tests over the Internet for delivery either via the web or by mail to me in the UK. Not only does this raise commercial questions regarding the 'exclusivity' of a local distribution agent's contract, it also raises questions about the conditions under which such materials are made available and the qualifications required of the user. By default, the current position

[1] This paper is a revision and update of an invited paper for a special edition of the European Journal of Psychological Assessment on Standards and Guidelines in Psychological Assessment. Sections of the original paper have been reproduced with permission from European Journal of Psychological Assessment, Vol 17 (3), 2001, pp. 173-186, © 2001 by Hogrefe & Huber Publishers, USA, Canada, Germany, Switzerland.

would seem to be that the user needs to be qualified in terms of the country of origin of the tests rather than the country in which they will be used.

The Internet is also making possible the development of extremely complex multi-national scenarios. Bartram (2000) presents the following example:
"An Italian job applicant is assessed at a test centre in France using an English language test. The test was developed in Australia by an international test publisher, but is running from an ISP located in Germany. The testing is being carried out for a Dutch-based subsidiary of a US multi-national. The position the person is applying for is as a manager in the Dutch company's Tokyo office. The report on the test results, which are held on the multi-national's Intranet server in the US, are sent to the applicant's potential line-manager in Japan having first been interpreted by the company's out-sourced HR consultancy in Belgium."

At present there are large variations between countries in the standards they adopt and the provision they make for training in test use. There are differences in the regulation of access to test materials and the statutory controls and powers of local psychological associations, test commissions and other professional bodies (Bartram & Coyne, 1998a). There is a clear need to establish international agreement on what constitutes good practice and what the criteria should be for qualifying people as test users.

International trends

A number of surveys have been carried out over the past few years that bear directly on international differences and similarities in patterns of test use and testing practice (Bartram, 1998; Bartram & Coyne, 1998a; Muniz, Prieto, Almeida & Bartram, 1999; Muñiz, Bartram, Evers, Boben, Matesic , Glabeke, Fernández-Hermida, & Zaal, 2001).

In 1996 the ITC and the European Federation of Psychologists Associations (EFPA) jointly devised a survey. Four language versions were produced (English, French, German and Spanish). The questions covered:

- Who uses tests?
- Availability of tests.
- Quality standards, codes and control mechanisms.
- Expertise and competence in test design, development and use.
- Opinions and beliefs about test use.

Complete data were received from 37 of the 48 countries in the sample. The design of the study was such that the responses were aggregated to form 'corporate' national responses representing the views of national psychologists' associations.

While a great deal of interesting information was obtained from this survey, of particular relevance to the present issue of guidelines on test use, it was found that:

- The majority of test users were estimated to be non-psychologists (86.3%).
    - The largest user group were educational (78.8%) while the smallest were clinical and forensic (11.8% and 0.4% respectively).
    - Nine percent of all test users were in the work area, of who about 65% were non-psychologists. This ratio varied quite a lot from country to country and is also thought likely to be an underestimate.
    - The clinical and legal areas are the only ones where non-psychologists do not outnumber psychologists as test users.
    - With reference to training and qualification in test use, it was found that:
    - Only 41% of users were estimated to have received any training in test use.
    - The lowest rates of training are in the educational testing area and the highest in clinical.
    - Psychologists only have slightly higher rates of training than non-psychologists. For example, in the work area, 54% of psychologists and 40% of non-psychologists are reported as having received training in testing.

Overall, the survey showed the need for more training for all test users (Bartram & Coyne, 1998a). However, it also revealed marked differences in patterns of response between different countries (Bartram & Coyne, 1998b).

While this survey provided a good view of the 'corporate' responses of the major national psychological associations, subsequent work by members of the EFPA Standing Committee on Tests and Testing has provided a more detailed view of the attitudes of individual

psychologists within a sample of European countries. In general, this showed that European psychologists have a very positive attitude towards tests and testing. They do, however, express the need for institutions to adopt a more active role in promoting good testing practices. The results show that the tests most frequently used by psychologists are Intelligence tests, Personality questionnaires and Depression scales. The patterns of use do however differ as a function of area of application (clinical, educational or work psychology) and country. A detailed report on this survey can be found in Muniz at al, (2001).

ITC Guidelines

The International Test Commission (ITC) is an "Association of national psychological associations, test commissions, publishers and other organisations committed to promoting effective testing and assessment policies and to the proper development, evaluation and uses of educational and psychological instruments." (ITC Directory, 2001). The ITC is responsible for the International Journal of Testing (published by Lawrence Erlbaum) and publishes a regular newsletter, Testing International (available from the ITC website). Three ITC projects bear directly on the theme of the present paper: the ITC Guidelines on Adapting Tests, the ITC Guidelines on Test Use and the ITC Guidelines on Computer-Based Testing and Testing on the Internet. All these guidelines can be obtained from the ITC website (www.intestcom.org).

Guidelines on Adapting Tests
These were developed by a 13-person committee representing a number of international organisations. The objective was to produce a detailed set of guidelines for adapting psychological and educational tests for use in various different linguistic and cultural contexts (Van de Vijver & Hambleton, 1996). This is an area of major importance as tests become used in more and more countries, and as tests developed in one country get translated or adapted for use in another. Adaptation needs to consider the whole cultural context within which a test is to be used. Indeed, the adaptation guidelines apply wherever tests are moved from one cultural setting to another - regardless of whether there is a need for translation. Hambleton (1994) describes the project in detail and outlines the 22 guidelines that have emerged from it. These guidelines fall into four main categories: those concerned with the cultural context, those concerned with the technicalities of instrument development and adaptation, those concerned with test administration, and those concerned with documentation and interpretation. All but the second of these also have direct implications for test use and for test users.

Guidelines on Test Use
The focus of this ITC project is on good test use and on encouraging best practice in psychological and educational testing. The work carried out by the ITC to promote good practice in test adaptations was an important step towards assuring uniformity in the quality of tests adapted for use across different cultures and languages. However, there are two key issues in psychological test practice. First, one has to ensure that the tests available meet the required minimum technical quality standards. Second, one needs to know that the people using them are competent to do so.

The Test Use guidelines project was started following a proposal from the present author to the ITC Council in 1995. The aim was to provide a common international framework from which specific local standards, codes of practice, qualifications, user registration criteria, etc could be developed to meet local needs. The intention was not to 'invent' new guidelines, but to draw together the common threads that run through existing guidelines, codes of practice, standards and other relevant documents, and to create a coherent structure within which they can be understood and used.

The competencies defined by the guidelines were to be specified in terms of assessable performance criteria, with general outline specifications of the evidence that people would need for documentation of competence as test users. These competences needed to cover such issues as:

- professional and ethical standards in testing,
- rights of the test candidate and other parties involved in the testing process,
- choice and evaluation of alternative tests,
- test administration, scoring and interpretation,
- report writing and feedback.

In the process of development (Bartram, 1998; Bartram, 2001), emphasis was placed on the need to consider and, where possible, consult a number of different stakeholders. These fall in to three broad categories:

1. Those concerned with the production and supply of tests (e.g. test authors, publishers and distributors)
2. The consumers of tests (e.g. Test users, test takers, employers and other third parties such as parents, guardians etc)
3. Those involved in the regulation of testing (e.g. professional bodies, both psychological associations and others, and legislators).

Just four years after the original proposal was put to the Council of the ITC, the ITC Council formally endorsed the International Guidelines on Test Use. During this period of four years, the Guidelines evolved from an initial framework document, through a series of workshops and consultation exercises into their present form.

The introduction to the Guidelines sets out who they are intended for and what other categories of people might find them of relevance. In addition to setting out a 'key purpose statement' for testing, detailed guidelines are provided on taking responsibility for ethical test use and on following good practice in the use of tests. The Guidelines also contain appendices dealing with:

- Organisational policies on testing.
- Developing contracts between parties involved in the testing process.
- Points to consider when making arrangements for testing people with disabilities or impairments.

The importance of 'context' has already been mentioned. The Guidelines, as written, are context-free. Guidelines reflect consensus on practice. They tend to be general and embody principles and have strong links to ethics through the process of defining what is meant by ‘good conduct’. In developing the ITC Guidelines a distinction was drawn between ‘good practice’ (what is expected of the competent practitioner) and ‘best practice’ (what is aspired to by many and attained by a few). The emphasis of the ITC Guidelines is on the former.

The Guidelines in Test Use project received backing from the BPS, APA, NCME, EAPA, EFPPA, and from a large number of European and US test publishers. Copies of the full Guidelines (in 14 different languages including English) can be obtained from the ITC website (www.intestcom.org) and were printed in the first edition of the ITC’s International Journal of Testing (ITC, 2001).

Guidelines on Computer-Based Testing and the Internet

In 2001, the ITC Council approved a project on developing guidelines on good practice for computer based and internet-based testing. The aim was not to ‘invent’ new guidelines but to draw together common themes that run through existing guidelines, codes of practice, standards, research papers and other sources, and to create a coherent structure within which these guidelines can be used and understood. Contributions to the guidelines have been made by psychological and educational testing specialists, including test designers, test developers, test publishers and test users drawn from a number of countries.

Furthermore, the aim is to focus on the development of Guidelines specific to CBT/Internet based testing, not to reiterate good practice issues in testing in general. Clearly, any form of testing and assessment should conform to good practice issues regardless of the method of presentation. These guidelines are intended to complement the existing ITC Guidelines on Test Use and on Test Adaptation, with a specific focus on CBT/Internet testing.

The first stage of this project involved a through review of existing guidelines and standards, especially those relating to computer-based testing. A small survey of test publishers was also carried out to identify key issues associated with testing over the Internet. This identified remote administration as a major issue for standards to address. The ITC held a conference in 2002 in Winchester, England that focused on the issues that Guidelines need to address. This was attended by over 250 experts from 21 different countries. All those who attended the conference and others on the ITC’s database were circulated with a first draft of the guidelines for comment. Detailed comments on the draft Guidelines were received from

individuals and organisations representing 8 different countries (Australia, Canada, Estonia, Holland, Slovenia, South Africa, UK and USA). This feedback together with material from the report of the APA Internet Task Force (Naglieri et al, 2004) was reviewed in detail and relevant points included within version (0.5) of the guidelines. This process was completed in February 2004.

A second consultation process was implemented in March 2004 (contacting the same individuals as before and using the ITC web site). Since then, there has been a further round of revisions, culminating in the latest revision, version 0.6. This, and other ITC Guidelines, can be found on the ITC's website (www.intestcom.org) .

In addition to formal consultations, input has been obtained through conferences and workshops around the world (e.g. UK, USA, Austria, and China). The Guidelines were approval by the ITC Council in July 2005.

The guidelines address four main issues:
1. Technology – ensuring that the technical aspects of CBT/Internet testing are considered, especially in relation to the hardware and software required to run the testing.
2. Quality – ensuring and assuring the quality of testing and test materials and ensuring good practice throughout the testing process.
3. Control – controlling the delivery of tests, test takertest-taker authentication and prior practice.
4. Security – security of the testing materials, privacy, data protection and confidentiality.

Each of these is considered from three perspectives in terms of the responsibilities of:
1. The test developer
2. The test publisher
3. The test user

A key feature of the Guidelines is the differentiation of four different modes of test administration:
1. Open mode – Where there is no direct human supervision of the assessment session. Internet-based tests without any requirement for registration can be considered an example of this mode of administration.
2. Controlled mode – Remote administration where the test is made available only to known test-takers. On the Internet tests, such tests require test-takers to obtain a logon username and password. These often are designed to operate on a one-time-only basis.
3. Supervised (Proctored) mode – Where there is a level of direct human supervision over test-taking conditions. For Internet testing this requires an administrator to log-in a candidate and confirm that the test had been properly administered and completed.
4. Managed mode – Where there is a high level of human supervision and control over the test-taking environment. In CBT testing this is normally achieved by the use of dedicated testing centres, where there is a high level of control over access, security, the qualification of test administration staff and the quality and technical specifications of the test equipment.

National and International Initiatives

Practitioners in the field want to insure that their practices conform to internationally recognised standards of good practice. However, it is not enough just to set standards. Having formulated standards, there is a need for independent examination to see whether in daily practice those standards are indeed met. In several places around the world, initiatives to set up independent quality audit procedures have been started.
In this Section some important current national and European initiatives are reviewed.

Developments in the USA

- The AERA/APA/NCME Test Standards

The various surveys carried out by the ITC and EFPPA have all shown that the AERA/APA/NCME (APA, 1985) Standards for Educational and Psychological Testing have been widely adopted as the authoritative definition of technical test standards. After a lengthy revision and consultation process, a new version of these influential Test Standards was published in 1999 (AERA, 1999). Many countries' psychological associations are likely to adopt these as the successor to the earlier edition. While the USA has provided a clear lead in this area, the issue of test user

qualification has not been addressed to the same degree as it has been in Europe. The position has tended to be one of assuming that those with a doctorate in psychology will have the necessary competence to be test users.

- APA Task Force on Test User Qualifications.

The APA Task Force on Test User Qualifications (TFTUQ) has developed guidelines that inform test users and the general public of the qualifications that the APA considers important for the competent and responsible use of psychological tests (DeMers et al, 2000). The term 'test user qualification' refers to the combination of knowledge, skills, abilities, training, experience that the APA considers optimal for psychological test use. In this sense, the word 'qualification' is being used to indicate competence rather than the award of some certificate or license or the outcome of a credentialing process.

The guidelines describe two areas of test user competence: (a) generic competences that serve as a basis for most of the typical uses of tests and (b) specific competences for the optimal use of tests in particular settings or for specific purposes. The guidelines provide very detailed discussions of competence requirements for a number of different testing contexts (e.g. healthcare, counselling, employment). The Guidelines are aspirational in that they define what the APA consider important for the optimal use of tests. As they make clear, the competences needed by any particular test user will depend on the use they will be making of tests and the context in which they will be doing testing. They further note that the testing process may be distributed between different individuals and make clear that the APA guidelines are directed primarily at the person who is responsible for the use of tests in the assessment process.

In relation to taxonomies of tests, the report discusses the three-level system (A, B, C) for classifying test user qualifications that was first defined in 1950 (APA, 1950). This system labelled some tests (Level A) as appropriate for administration and interpretation by non-psychologists; others (Level B) as requiring "some technical knowledge of test construction and use, and of supporting psychological and educational subjects such as statistics, individual differences, the psychology of adjustment, personnel psychology, and guidance"; and others (Level C) as being restricted to "persons with at least a Master's degree in psychology, who have had at least one year of supervised experience under a psychologist".

While the APA dropped this classification from the 1974 and subsequent editions of the Standards for Educational and Psychological Testing, it has remained in use by many test publishers. The Task Force's report suggests that this method of classification is now obsolete and needs to be replaced by one where the competences required are defined in relation to the context, instrument and use to which it will be put.

These Guidelines were approved by the APA Council in August 2000. The Task Force are now considering ways in which they might best be disseminated.

- The ATP Guidelines on Computer-based testing.

In the area of computer-based testing, the Association of Test Publishers (ATP, 2002) has developed technology-based guidelines for the testing industry. Though the ATP is primarily US-focused in terms of its membership and the services it provides to its members, it does have most of the major international publishers as members. Given the arguments presented earlier about the internationalisation of testing, the work of the ATP should be of wide interest. Indeed, the goal for the proposed guidelines will be international adoption by companies involved in technology-based testing.

The guidelines are intended for use as an aid in the development, delivery and use of computer-based certification examinations as well as aptitude testing in general. The ATP want to see the guidelines used to aid in the development, delivery and publishing of technology-based tests and assessment instruments. The guidelines cover applications for use on the Internet and various multimedia computer strategies used to deliver, administer and score tests. Work on these guidelines identified a number of issues that have not been adequately addressed by existing

standards and guidelines. Some of these issues include the development of standards on immediate score reporting, item banking and models for estimating parameters. New types of response models need new standards that everyone can use.

- APA Task Force on Testing on the Internet

Recently, this APA Task Force has issued a report (Naglieri et al, 2004) discussing a range of issues related to Internet Testing. While not a set of standards of guidelines, the Task Force has provided some guidance on how ethical standards and good practice relate to Internet Testing. This work has been incorporated into the ITc Guidelines on Computer-Based Testing and Testing on the Internet.

- Test Taker Rights and Responsibilities: Working Group of the Joint Committee on Testing Practices

This Joint Committee produced a useful set of guidelines (Joint Committee on Testing Practices, 2000) which focus on test takers. These consider testing from the point of view of the person who is being tested and tries to identify what their rights and responsibilities are in the testing process. This is an important area, often missed in other standards.

National and Regional Developments in Europe

- The BPS test user competence approach

In the UK the British Psychological Society (BPS) fulfils the role of quality auditor by setting test use standards, defining test review criteria and accrediting those who will assess the competence of test users (Bartram, 1995, 1996). The approach adopted in Britain over the past fifteen years has specifically addressed the issue of how to implement standards of competence as deliverable outcomes. The BPS Steering Committee on Test Standards (SCTS) developed a strategy for combating both the problems of poor tests and bad test use by focusing on the would-be test user. The overall strategy was to develop more competent test users and to provide better information for them about tests. The latter has been accomplished through the establishment of a Register of Test Users with its associated journal Selection and Development Review; and the publication of regular test reviews and updates (Bartram et al, 1992, 1995; Lindley et al, 2000).

To accredit test user competence, a certification process was implemented in 1991. This had been confined to psychological test use in occupational assessment settings, but a new test user certificate covering educational testing was launched in 2004. The occupational test user qualification comprises a number of certificates covering test administration, basic psychometric principles and the use of tests of ability and aptitude, and the use of more complex instruments, particularly those used in personality assessment. Details of the background to the development of the BPS approach and to the progress made in its implementation are given in Bartram (1995, 1996).

There has been a large take up of the certificate. A total of over 18,000 people have obtained the Level A qualification with over 6000 having the Level B one (the latter has been available for a shorter period of time than the former). This has provided the BPS with sufficient income to ensure that the process of maintaining high standards can be adequately resourced. There is now a wide acceptance within the professional community (not just among psychologists), that the BPS accredited qualifications represent the yardstick by which user competence is judged. The qualifications have tremendous currency and have entered the language as a short-hand ways of referring to particular levels of expertise.

In January 2003, the BPS formally launched The Psychological Testing Centre (PTC). This is a single body responsible within the Society for all issues to do with the testing. The PTC acts as the interface with the various stakeholders in testing (users, test takers, publishers, trainers and so on) and is responsible for the management and delivery of qualification, test reviews and so on. The PTC manages its own informative website (www.psychtesting.org.uk) from which a wide range of information and documents are available. Full details of the BPS Test user standards and qualifications can be downloaded from this site. It also provides access to all the test reviews (for a charge).

- Swedish Foundation for Applied Psychology (STP)

In Sweden an independent institute (the STP), similar in concept to the PTC, provides test user qualification certification, test reviewing, the quality audit of organisational testing policies and an ombudsman function for test takers to appeal to. Originally established in 1966 by the Swedish Psychological Association, new goals were set for the STP in 1996. The STP is an independent, non-profit organisation working to obtain general consensus amongst the key stakeholder groups on quality in tests and test use. It provides an independent forum for professional test users, universities, professional associations, developers and publishers and has developed a quality model using a network of representatives of all stakeholder groups.

Test and test use quality assurance build on the ITC Guidelines on Test Use (which have been translated into Swedish for the STP). STP also publishes test reviews based on the BPS model and has developed a scheme for the certification of test-user competence following the BPS model. The STP also accredits organisational test policies and testing processes. Guidelines for organisational policies on test use are based on the model provided in the ITC Guidelines on Test Use.

More recently, STP is working closely with the Norwegian Psychological Association (NPA). The NPA together with Det Norske Veritas (DNV) is established a quality assured procedure for the delivery of test user qualification. Both the NPA and STP are planning to use DNV as their independent quality assurance body. DNV specialise in the delivery of a range of ISO standards-based procedures.

- The Institute for Best Test Practice in the Netherlands

The Institute for Best Test Practice in The Netherlands was set up with a very similar set of aims and objective to the Swedish STP - though the two organisations have some differences in approach. Like the STP, the Dutch Institute has taken the ITC Guidelines on Test Use as a basis for building the standards it will promulgate, thus ensuring that its approach will be consistent with the International consensus on what good practice is. The Institute has now become a part of CITO.

In addition to stressing the need for quality, the objectives of the Institute are for transparency. First, the standards adopted must be open and available for all to see. Second, all the results of examinations carried out by the Institute will be published on the Internet: Registers of certified test users; reviewed tests; and quality audited organisations. Everybody interested in the quality of people, instruments and practices will therefore be able to access this information. In this way, it is hoped that the new Institute will provide a practical way of implementing quality and transparency.
Both the Dutch and Swedish developments are quite new.

- Guidelines for the Assessment Process

The European Association of Psychological Assessment set up a Task Force under the direction of Prof Rocio Fernandez-Ballesteros to develop Guidelines for the Assessment Process (GAP: Fernandez-Ballesteros, 1997; Ballesteros, et al., 2001). The results of this work are a set of guidelines that look broadly at assessment as a process, particularly as a clinical intervention, rather than just focusing on tests and testing.

- The German DIN 33430 project

In Germany work has also been done on looking at assessment as a process. The focus of this is rather different to the GAP project. Instead the focus is specifically on assessment for selection and recruitment. Ackerschott (2000) reported that 80% to 90% of psychological assessment services are sold in Germany by non-psychologists with a wide range of different backgrounds in terms of skill and experience. He notes that there is currently no control or regulation regarding the quality of the services they provide. All sorts of different tests are used in the assessment process. The test-commission of the Federation of German Psychologists Associations has had no visible impact on this situation. Because of concerns over this situation, in 1995 the BDP officially initiated the DIN 33430 project by applying to the German

Association of Standardization for a standard of psychological tests. The objectives of this BDP-initiative were to:

- protect candidates from unprofessional use or misuse of tests and assessment-procedures;
- minimise wrong decisions in the context of aptitude-testing and the subsequent economic, social and personal costs;
- require test-developers and publishers to raise the quality of tests;
- encourage good practice in the implementation of psychological assessment-procedures, tests and other psychological instruments.

The German DIN 33430 project has, after consultations and discussions with diverse groups focused on the following as its subject: Requirements for Procedures / Methods and their Applications in the Context of Judgements of Professional Aptitude (generally for selection, either internal or external). The DIN 33430 has now been published by the German Standards Institute (DIN) and has the role of a non-mandatory set of guidelines for good practice. Organizations which show their procedures meet this standard can be DIN 33430 accredited, in the same way as organisations can achieve ISO 9000 standards.

- The ISO PC230 project

In 2007 the International Standards Organization (ISO) started a project to set a standard for assessment in work and organizational settings. This was initiated by the DIN and their work on DIN 33430. Work on the ISO standard is currently progressing and is expected to result in a service delivery standard for defining quality in the delivery of assessment service sin client contractor relationships.

The work of the EFPA Standing Committee on Tests and Testing

The work of the EFPA Standing Committee in gathering information through surveys has already been reviewed. A further and potentially more far-reaching initiative being pursued by the Committee is that of setting a common set of European criteria for test reviews and for test user competence.

The EFPA Test Review Criteria

There are currently two well-established test review procedures in place in Europe: the Dutch COTAN process, and the British BPS test review process. The Spanish also developed a process that drew from both the Dutch and British experience. In Sweden, the STP adopted a approach very similar to that used in Britain.

The first stage of developing a European framework for test reviewing was to combine the best features of the British, Dutch and Spanish procedures into a single document. This and the detailed test review criteria were reviewed and sent out for consultation across Europe, and were adopted by the EFPA Committee in 2001. The EFPA Test Review Criteria and supporting documentation are available from the EFPA website: www.efpa.be.

Since their acceptance by EFPA, the British Psychological Society has adopted these standards as the basis for all its new reviews and has been updating existing reviews to fit the new format and structure. All the British reviews are now available online on the PTC website: www.psychtesting.org.uk.

European Test User Standards and national certification systems

More recently, EFPA in conjunction with the European Association of Work and Organizational Psychologists (EAWOP) has started work on developing a European set of standards defining test user competence. This project is running in parallel with a major review in the UK of the BPS standards for test use and work in Sweden, Norway and Denmark on the development of competence-based test user certification procedures.

One spin off from this work was the development of a set of standards for occupational assessment. These are now being incorporated into the ISO PC230 standard.

For all these projects, the ITC Guidelines on Test Use are being used as the framework for the standards. Where testing relates to computer-based delivery – which is increasingly the case these days – the ITC Guidelines on computer-based delivery can be consulted.

Other standards specifically relating to computer-based assessment

Valenti et al (2002) reviewed use of ISO 9126 as a basis for CBA system evaluation. ISO9126 is a standard for Information Technology – Software Quality characteristics and sub characteristics. The standard focuses on: Functionality; Usability; Reliability; Efficiency; Portability and Maintainability. Valenti et al (2002) base their review around the first three of these.

The British Standards Institute published a standard (BS7988) in 2002: A Code of Practice for the use of information technology for the delivery of assessments. The Standard relates to the use of Information Technology to deliver assessments to candidates and to record and score their responses. The Scope is defined in terms of three dimensions - the types of assessment to which it applies, the stages of the assessment 'life cycle' to which it applies and the Standard's focus on specifically IT aspects.

This standard has now been incorporated into an ISO standard: Information technology -- A code of practice for the use of information technology (IT) in the delivery of assessments. [Educational] (ISO/IEC 23988: 2007). ISO23988 is designed to provide a means of:

- showing that the delivery and scoring of the assessment are fair and do not disadvantage some groups of candidates, for example those who are not IT literate;
- showing that a summative assessment has been conducted under secure conditions and is the authentic work of the candidate;
- showing that the validity of the assessment is not compromised by IT delivery; providing evidence of the security of the assessment, which can be presented to regulatory and funding organizations (including regulatory bodies in education and training, in industry or in financial services);
- establishing a consistent approach to the regulations for delivery, which should be of benefit to assessment centres who deal with more than one assessment distributer;
- giving an assurance of quality to purchasers of "off-the-shelf" assessment software.

It gives recommendations on the use of IT to deliver assessments to candidates and to record and score their responses. Its scope is defined in terms of three dimensions: the types of assessment to which it applies, the stages of the assessment "life cycle" to which it applies and its focus on specifically IT aspects. The scope does not include many areas of occupational and health related assessment. While it includes “Assessments of knowledge, understanding and skills (i.e. achievement tests)” it excludes “psychological tests of aptitude and personality”

Conclusions

A great deal of progress has been made in the development and dissemination of standards and guidelines to help improve testing and test use. The ITC Guidelines have rapidly become accepted as defining the international framework within which local standards should fit, and onto which they should be mapped. Different countries have explored and are exploring ways of delivering quality, through test reviewing and registration, test user training, accreditation procedures, and the establishment of test institutes.

Perhaps the greatest challenge is that of test 'benchmarking'. It is argued by many that both test takers and test users need to have some form of quality stamp on a test to indicate that it meets minimum 'safety' standards. This need has been particularly strongly expressed in relation to Internet testing. Test users, test takers and test publishers appear to be in broad agreement on the need for some way of awarding tests that meet minimum technical standards some form of 'stamp of approval' to differentiate them from the mass of instruments and questionnaires for which no psychometric data are available. The BPS will be introducing a test registration system in Britain in 2005. This will be based on the standards defined in the EFPA Test Review Criteria, and will provide publishers and developers of tests with a ‘quality stamp’ that allows them to differentiate genuine psychometric tests from other less rigorously developed instruments.

What we can be sure of is that the growth of the Internet as a medium for the delivery of tests 'at a distance' will increasingly impact on testing. It has already raised a host of issues that we need to address in relation to good practice. The ITC, BSI, ISO, ATP and APA have all picked up on these issues and we should look forward to the evolution of clearer standards and guidelines for best practice in

this area as this technology matures. The ITC will continue to pick up on these national and regional developments and take forward its role of providing an internationally agreed framework within which local diversity can be accommodated.

References

Ackerschott, H. (2000). Standards in Testing and Assessment: DIN 33430. Paper presented at the Second International Congress on Licensure, Certification and Credentialing of Psychologists, OSLO, Norway

APA (1985). American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). Standards for Educational and Psychological Testing. Washington DC: American Psychological Association.

AERA (1999). American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. Standards for Educational and Psychological Testing. Washington DC: American Educational Research Association.

APA (1950). Ethical standards for the distribution of psychological tests and diagnostic aids. American Psychologist, 5, 620–626.

ATP (2002) The ATP Guidelines on Computer Based Testing. Association of Test Publishers.

Bartram, D., & Coyne, I. (1998a). The ITC/EFPPA survey of testing and test use within Europe. In Proceedings of the British Psychological Society's Occupational Psychology Conference (pp. 197-201). Leicester, UK: British Psychological Society.

Bartram, D., & Coyne, I. (1998b). Variations in national patterns of testing and test use. European Journal of Psychological Assessment, 14, 249-260.

Bartram, D., Lindley, P.A., & Foster, J. (1992) Review of Psychometric Tests for Assessment in Vocational Training. Leicester, England: BPS Books.

Bartram, D. (Senior Editor) with Anderson, N., Kellett, D., Lindley, P.A. and Robertson, I. (Consulting Editors). (1995). Review of Personality Assessment Instruments (Level B) for use in occupational settings. Leicester: BPS Books.

Bartram, D. (1995). The development of standards for the use of psychological tests in occupational settings: The competence approach. The Psychologist, May, 219-223.

Bartram, D. (1996). Test qualifications and test use in the UK: The competence approach. European Journal of Psychological Assessment, 12, 62-71.

Bartram, D. (1998). The need for international guidelines on standards for test use: A review of European and International initiatives. European Psychologist, 3, 155-163.

Bartram, D. (2000). Internet recruitment and selection: Kissing frogs to find princes. International Journal of Selection and Assessment, 8, 261-274.

Bartram, D. (2001) The development of international guidelines on test use: the International Test Commission Project. International Journal of Testing, 1, 33-53

Fernandez-Ballesteros, R. (1997). Task force for the development of guidelines for the assessment process (GAP). The International Test Commission Newsletter, 7, 16-20.

Fernandez-Ballesteros, R., De Bruyn, E.E.J., Godoy, A., Hornke, L.F., Ter Laak, J. Vizcarro, C., Westhoff, K., Westmeyer, H., & Zaccagnini, J.L. (2001). Guidelines for the assessment process (GAP). European Journal of Psychological Assessment, 17, 187-200.

DeMers, S.Y., Turner, S.M. (Cochairs), Andberg, M. Foote, W. Hough, L. Ivnik, R. Meier, S. Moreland, K. & Rey-Casserly, C.M. (2000). Report of the Task Force on Test User Qualifications. Washington, D.C.: Practice and Science Directorates, American Psychological Association

Hambleton, R. (1994). Guidelines for adapting educational and psychological tests: A progress report. European Journal of Psychological Assessment, 10, 229-244.

ITC (2001). International Test Commission Directory. http://www.intestcom.org

ITC (2001). International Guidelines on Test Use. International Journal of Testing, 1, 95-114.

Joint Committee on Testing Practices. (2000). Rights and Responsibilities of Test Takers: Guidelines and Expectations. Washington DC: Joint Committee on Testing Practices.

Lindley, P.A. (Senior Editor) with Banerji, N., Cooper, J., Drakeley, R., Smith, J.M., Robertson, I., & Waters, S. (Consulting Editors) (2000). Review of personality assessment instruments (Level B) for use in occupational settings. 2nd Edition. Leicester: BPS Books.

Muñiz, J., Prieto, G., Almeida, L., & Bartram, D. (1999). Test use in Spain, Portugal and Latin American countries. European Journal of Psychological Assessment, 15, 151-157.

Muñiz, J., Bartram, D., Evers, A., Boben, D., Matesic, K., Glabeke, K., Fernández-Hermida, J.R., & Zaal, J.N. (2001). Testing Practices in European Countries. European Journal of Psychological Assessment, 17, 201-211.

Naglieri, J.A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., & Velasquez, R. (2004). American Psychologist, February, 2004.

Valenti, S., Cucchiarelli, A., & Panti, M. (2002). Computer based assessment systems evalution via the ISO9126 quality model. Journal of Information Technology Education, 1(3), 157-175.

Van de Vijver, F. & Hambleton, R. (1996). Translating tests: some practical guidelines. European Psychologist , 1, 89-99.

Important websites

- The American Psychological Association: www.apa.org
- The British Psychological Society's Psychological Testing Centre: www.psychtesting.org.uk
- The European Federation of Psychologists Associations: www.efpa.eu
- The International Standards Organization: www.iso.org
- The International Test Commission: www.intestcom.org

The author:

Prof Dave Bartram
Research Director
SHL Group plc
The Pavilion, 1 Atwell Place. Thames Ditton,
Surrey, UK, KT7 0NE

E-Mail: Dave.Bartram@SHLGroup.com

WWW:
http://www.shlgroup.com

Prior to joining SHL in 1998, Dave Bartram was Dean of the Faculty of Science and the Environment, and Professor of Psychology in the Department of Psychology at the University of Hull. He is Past-President and a Council member of the International Test Commission (ITC), a past Chair of the British Psychological Society's Steering Committee on Test Standards and Chair of the European Federation of Psychologists Association's (EFPA) Standing Committee on Tests and Testing. He is President of the International Association of Applied Psychology's Division 2 (Measurement and Assessment). He led the development of the ITC's International Guidelines for Test Use and, with Iain Coyne, the development of the ITC Guidelines for Computer-based and Internet Delivered Testing. He is currently a member of an ISO project committee developing an international standard for assessment in work and organizational settings. He is the author of several hundred scientific journal articles, papers in conference proceedings, books and book chapters in a range of areas relating to occupational psychology and assessment, especially in relation to computer-based testing.

He received the award for Distinguished Contribution to Professional Psychology from the BPS in 2004 and was appointed Special Professor in Occupational Psychology and Measurement at the University of Nottingham in 2007.

Examinations in Dutch secondary education - Experiences with CitoTester as a platform for Computer-based testing

Mark J. Martinot
Cito, Unit Secondary Education

Abstract:
This article concerns the introduction of ICT in the national final examinations for Dutch secondary education. The use of ICT in examinations could add surplus value in many areas. This includes the area of test content, logistics and psychometrics. Various pilot projects have been carried out in recent years to gain experience with administering examinations by computer, rather than on paper. These pilot projects provide a platform for studying substantive aspects of examinations (added value), technical aspects (software and system requirements) and organizational aspects (examination and administration process). The participating schools make use of CitoTester, the test program developed by Cito for administering computer-based examinations. In general, schools have reacted favourably. In the coming years, the number of computer-based exams will grow. There will be additional investments in a national standard for computer-based examinations, in which conditions relating to system requirements, installation, interfaces and user potential of examination software are well defined.

Introduction

The Cito organization is expert in the field of valid, reliable measurement of learning performance. On instruction from the government, Cito develops the national examinations in the Netherlands for preparatory intermediate vocational education, higher general secondary education and pre-university education. As an expertise centre, Cito also does research in and offers advice for modernizing national examinations.

The examinations in the Netherlands are the responsibility of the minister of Education, Culture and Science. Various parties collaborate on creating the examinations. As stated above, Cito is responsible for designing examination questions. The Cevo, a national committee set up by the minister, is responsible for the examination process, the specification of test-designs, and the approval of the questions to be used in the examinations. The IB Group is an independent administrative body that, on instruction from the minister, is responsible for the logistics of the examination process.

Annually, at the end of secondary education, some 200,000 students in 700 schools take part in the national examinations. Each year, Cito designs more than 500 different tests for all subjects of the various types of education. In most cases the questions are presented to students on paper. However, more and more opportunities arise to administer examinations by computer instead of on paper. Schools are acquiring adequate ICT infrastructure and related knowledge to use this. Computers play an increasingly important role in education. And outside the schools as well, students are making use of ICT and computers in a growing number of situations. Together with several schools, Cevo, Cito and IB Group have therefore started various pilot projects to gain experience in the use of ICT in examinations.

ICT and examinations

The potential of ICT offers an alluring prospect. Compared to examinations on paper, the use of ICT in examinations could generate added value in many areas. Computer-based examinations can incorporate new elements, such as audio and video fragments, hot spots and drag-and-drop actions (clicking on and moving objects), and even simulations. This allows testing of other skills and can make examinations more attractive to students. At the same time, such examinations require fewer language skills than examinations on paper - something that can be beneficial especially for preparatory intermediate vocational students. Naturally, ICT can also ensure that many of the students' answers are checked automatically. Further computerization of the examination process enables a more flexible organization of national examinations. Schools can then offer more examination-moments, or organize the

examinations in different periods than is the case in the current situation. The use of digital examinations provides many changes with respect to those given on paper. The software has to meet various conditions. Below is a brief discussion of several aspects relating to test content, logistics and psychometrics.

The introduction of digital testing raises all kinds of questions relating to the quality of a computer based test as a measuring instrument. It is by no means self-evident that questions offered by the computer relate to the same skills of students as do questions presented on paper. In other words, thought must be given to the types of questions and assignments presented to students by computer, and to the desired interactions between student and computer. And there are choices to be made concerning the composition of tests from individual questions (e.g. linear or adaptive).

Additionally, careful attention has to be given to the reliability and the security of the testing system. With high-stake tests, such as national examinations, taking the test is a tense moment for students. For this reason, it is important that students are enabled to complete a computer-based exam without problems, even in the case of technical malfunction such as a computer failure or a poor Internet connection. It is also very important that the confidentiality of the test-questions be guaranteed. In no case should it be possible that unauthorized users have access to the tests or the testing system. And of course, there are various requirements that have to be met with respect to the security and quality of the transport and storage of data.
Finally, the handling of data made available through computer-based tests is a special point of attention. Large-scale administering of tests produces many test data that require further analysis. In many cases, this concerns advanced psychometric analyses involving the use of special software. An interactive question in a computer-based test can produce more - and different - data than a traditional question in a paper-based test. It is expected, therefore, that analysing software will have to be adapted in order to be able to process data from new types of questions. Proper agreements must be made about standards for classifying and exchanging these data, in order to provide for a proper connection between the testing software and separate analysis programs.

Cito Tester

To make it possible to offer computer-based examinations, Cito has started some time ago with the development of test software that meets the requirements and procedures for national examinations in secondary education. It was decided to introduce a modular design so that the various components could be adapted and maintained independently of one another.

In recent years, initial versions of this software have become available under the name of CitoTester. They were tested in pilot projects with schools in preparatory intermediate vocational education, intermediate secondary education and pre-university education. These pilots were so successful that for many subjects (especially in preparatory intermediate vocational education) CitoTester examinations are already being administered on a larger scale.

Below is a brief explanation of the design and possibilities of the CitoTester program. The software is still under development. New versions are released to meet the wishes of schools that arise from evaluations of the pilot projects. Expectations are that future versions of the software will be issued under the name ExaminationTester.

The current version of CitoTester comprises the following modules: TestManager, TestCenter and MarkingManager. TestManager is the module used by the test supervisor to manage the digital tests and to give students and teachers/markers access to these tests. TestCenter is the module that the students use to log in and take examinations. MarkingManager is the module that teachers/markers use to check and mark students' answers to open-ended questions.

The modules require a single installation on the computers of the network (server and client PCs) of the school or other official test administration location. The local server then exchanges data with the central server of Cito on the Internet. No working Internet connections are needed when tests are administered: all data are available both on the local server and on the central server of Cito.

The CitoTester software itself does not contain any examinations or examination questions.

Each digital examination is provided to a school as a separate file (examination package), for example on CD-ROM or via a secure Internet connection. The examination packages are secured with encryption and delivered with a time lock. They can only be installed in CitoTester by authorized users.

The images contain examples of questions that could be presented to students in TestCenter.

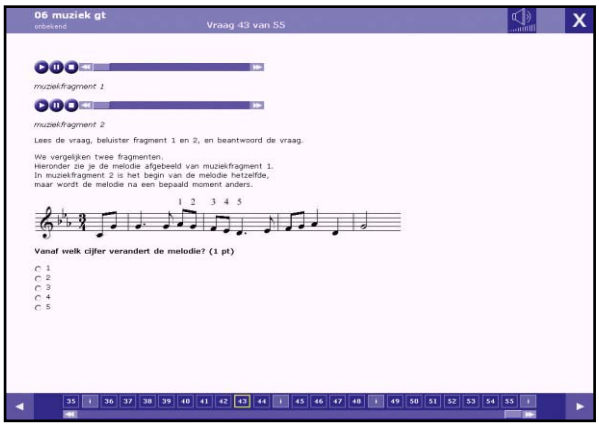


Figure 1: Sample music question

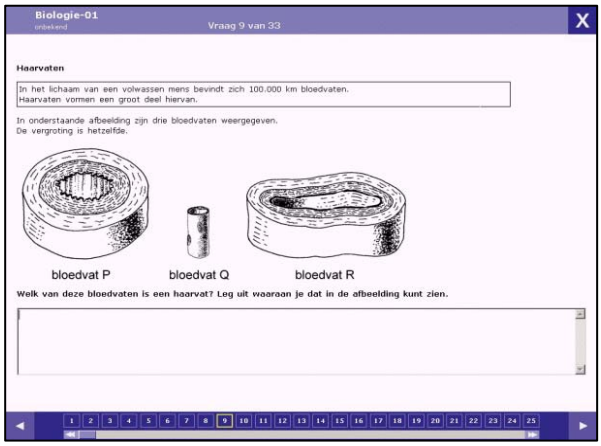


Figure 2: Sample biology question

At the bottom of each question window, a navigation bar is visible. Students can jump to each question by clicking on the correct question number in the navigation bar. Using the arrows at the left and right sides of this bar, students can also easily view the previous or next question. The bar at the top of the screen is the title bar. It contains the name of the exam that is being administered and the name of the student who is taking the test.

The question itself is in the centre of the screen. This is a question or an assignment for the student, possibly with contextual or background material, such as the audio fragments in Figure 1 or the drawings of a blood vessel in Figure 2. Various types of questions can be offered in the CitoTester environment. Questions could also contain extensive simulations in which students are required to perform various actions.

The screen layout of the questions in TestCenter applies to all examinations. Once they get used to the interface, students can take part in new digital examinations without supplementary instructions.

The TestCenter-module does not allow the student to use Windows and Internet functionality (e.g. email, chat) during the test. The answers given by the students will be stored securely in a central data folder. In case of a computer malfunction during an examination, the student can simply switch to another computer and continue, without losing any of the previous answers.

The school normally decides which students have access to specific examinations and at which times. In CitoTester this can be arranged with the TestManager module. The module is menu-driven; the main menu options appear directly after starting the TestManager (see Figure 3).

Figure 3: Test Manager menu

With Test Manager, schools can view and change the particulars of students, plan examination dates and retrieve various reports with examination results. If an examination fully consists of questions that can be marked automatically, the results are immediately available. Examinations with open-ended questions must first be marked by a teacher or an independent marker.

The teachers/markers concerned use the MarkingManager module for this purpose. This module supports the single marking of answers by one teacher/marker, and any supplementary multiple marking by one or more independent markers. The MarkingManager can be used on any Internet computer. Teachers/markers can therefore do their work at school or at home. If desired, Test Manager can automatically convert scores to grades. The test data are sent to Cito automatically.

Implementation

Before computer-based examinations become available nation-wide, pilot projects are carried out to gain the necessary experience with the software and (prototypes of) computer-based examinations.

These pilot projects provide a platform for studying substantive aspects of the examinations (the added value of ICT for specific subjects), technical aspects (software and system requirements) and organizational aspects (examination and administration process). To collect opinions and experiences, question lists are provided to school heads, teachers and students.

The pilot projects go through various stages. In the first instance, there is a 'proof of concept', which tests prototypes of digital examinations. This is done to see whether the technical solutions work well in practice and whether there is sufficient support among the parties concerned for the chosen design. Following this, in a limited number of schools, a trial test is given to students in a small-scale pilot project. During this stage it will become clear whether the way in which examinations at the pilot schools are organized and administered goes as expected. This test will also deliver the initial student data for statistical analysis. A continuation path will then commence, in which larger numbers of schools can participate. During a specific period, schools will have an opportunity to opt for computer-based examinations. Afterwards, it will be decided whether computer-based examinations should be introduced nationally.

In general, the reactions of schools have been favourable. Teachers say they appreciate the use of ICT, especially in cases where it results in a better fit between the examinations and the subjects taught, including the use of more realistic contexts in the questions. Students also favour computer-based examinations. They see the clear test structure provided by computers as a benefit (only one question at a time appears on the display). In most cases they also find the design of the questions on computers more appealing than that on paper. However, students also say that they consider it important to have sufficient practice in advance. ‘Practice examinations’ will therefore be sent to schools well in advance of the examination period. Experiences so far clearly indicate the need for offering adequate support to schools that use computer-based examinations for the first time. In practice, the need for such support declines sharply over time.

In the future, schools will gain experience for an increasing number of subjects with the aforementioned design of computer-based examinations. There will be new pilot projects with new versions of CitoTester or ExaminationTester. The results will affect the continued development of computer use for national examinations in the Netherlands. In addition, it will also be necessary to invest in (national) standards that will apply to computer testing. It is important that all parties concerned - schools, students and the parties involved in creating the examinations - agree on a generic solution, one that applies (in principle) to all computer-based examinations, in which the preconditions relating to system requirements, installation, interfaces and the user potential of the examination software are well defined.

The author:

Mark. J. Martinot
Cito, Unit Secondary Education
P.O. Box 1034
6801 MG Arnhem
The Netherlands
E-Mail: mark.martinot@cito.nl

M.J. Martinot, project manager of ICT and examinations at Cito, is responsible for developing and administering experimental computer-based examinations for Dutch secondary education. Related points of attention include devising suitable examination questions and context material for presentation on the computer, and establishing the desired functionality of the software required to administer computer-based examinations on a large scale.

Quality criteria in Open Source software for computer-based assessment

Dipl.-Psych. Annika Milbradt
RWTH Aachen University, Department of Industrial and Organizational Psychology

Abstract
The article presents quality criteria for software in general and discusses their importance for software used in computer-based assessments. On the example of testMaker, an Open Source software developed for web-based Self-Assessments, means to ensure software quality are illustrated. Also, the specifics and added value of Open Source software are commented on.

When planning an assessment project, ensuring quality of the assessment is something that one needs to think about at the very beginning. This is especially true if one aims at conducting a computer-based assessment. In this case, two different aspects influence quality: the content of the assessment and the software to be able to realize it. In order to define quality criteria either the individual aspects or their concurrence can be considered. It is not surprising, though, that existing guidelines on quality criteria resemble this fact. They focus on the *assessment* itself, such as the German norm for psychological aptitude diagnostics "DIN 33430 - Requirements for procedures and their application in job related proficiency assessment" (DIN Deutsches Institut für Normung e.V., 2002; see Westhoff et al., 2004, for an English version), on the *software* like the ISO/IEC 9126 Software engineering (International Organization for Standardization, 2001-2004) or on the *concurrence of assessment and software*, e.g. the "International Guidelines on Computer-Based and Internet Delivered Testing" (International Test Commission, 2005).

While content is by far the more important part of the assessment, this article will mainly be concerned with the software. Although software is basically just a means to an end in the assessment process, it is a necessary condition for a computer-based assessment to work. It is not hard to imagine that quality aspects of the software affect the assessment significantly. For example, if due to a software error the data collected in an assessment were not saved, the whole assessment would be to no avail. Consequences aren't always that dramatic, but it makes clear that quality of software is a factor that mustn't be ignored.

Quality criteria of software

One of the most common standards for the evaluation of software quality is the international norm ISO/IEC 9126. It consists of four parts of which the first one defines six major quality criteria of software. The first criterion is called *Functionality* and considers to what extent the software offers the required functions and their specified properties to satisfy stated or implied needs. The second criterion, *Reliability*, refers to the question if the software is capable to maintain its level of performance under stated conditions for a stated period of time. Directly related to human-machine-interaction and therefore especially meaningful in psychological assessments is the third criterion, *Usability*. It is concerned with the effort needed for use as well as the individual judgment of such use by a stated or implied set of users. The fourth criterion, *Efficiency*, takes a closer look at the relationship between the level of performance of the software and the amount of resources that are used under stated conditions. *Maintainability* as the fifth criterion is concerned with the effort needed to make specified modifications. And finally the sixth criterion, *Portability*, observes the ability of software to be transferred from one environment to another.

All of these criteria have sub-characteristics which again are divided into attributes. Although these attributes explain how to measure quality, the norm and its components remain abstract concepts and their meaning for the implementation of specific software needs to be interpreted in practice. The fundamental question is how software quality in terms of the named criteria can be ensured. A theoretical and general answer to that is taking these

quality criteria into consideration in the design and the development of the software and conducting tests to evaluate the software. A more practical and nongeneric answer shall be given on the example of the Open Source software testMaker (Milbradt, Zimmerhofer & Hornke, 2007) in the following.

A software for web-based Self-Assessments

The software testMaker has been developed for the implementation of web-based Self-Assessments. This term refers to self-directed counseling tools that address high school students attending grade 11 or higher. Self-Assessments consist of different tests and questionnaires related to the requirements of a specific field of study. They offer the opportunity to explore one‘s own abilities and academic orientations and include an articulate feedback about the participant‘s individual strengths and weaknesses. Especially in Germany, they have become a popular tool used by universities in order to counsel, select and bond prospective students (Zimmerhofer, Heukamp & Hornke, 2006), but they exist in other countries as well. In 2002, the Self-Assessment for Electrical Engineering and Computer Engineering (Zimmerhofer, 2003; Weber, 2003) has been the first one of its kind at RWTH Aachen University and one of the first ones in Germany. Today, several projects have followed including a Self-Assessment for international students interested in a technical field of study in Germany (available in German and English, see http://www.self-assessment.tu9.de). In order to participate in one of the Self-Assessments, a student visits the website and has to register first. Then, the Self-Assessment starts with an address of welcome explaining the goals and benefits of the tool followed by questions about demographic characteristics and educational and vocational plans. Afterwards, different tests, e.g. measuring verbal and figural reasoning, and questionnaires, e.g. measuring interest, motivation or self-efficacy, are provided to the participant. After completion, an automated feedback is given containing presentation and interpretation of the personal results. The Self-Assessment concludes with evaluation questions about the perceived benefit and acceptance and finally further information about student guidance.
To be able to administer the Self-Assessments with all of their components on the Internet, a technical solution was needed. Hence simultaneously to the content development of the Self-Assessment, a set of requirements towards software was defined. Listing all of the requirements would exceed the length of this article, but a very short excerpt shall be given. For example, related to the creation of questions (items) it was necessary to realize different item formats, integrate media (graphics, audio, video), limit the work time in achievement items, present the items in a standardized look to all participants and give the possibility to carry out adaptive tests for a short total duration of the Self-Assessment. Requirements connected to the scoring of data were automated individual feedback for participants (containing score, percentage and percentile rank), convenient data administration and data export or a personalized certificate of attendance. Whereas some of them demands hold true for online surveys in general, some of these are specific for assessments (like the limitation of work time) or are even specific for Self-Assessments (like the feedback options) (see Milbradt & Putz, 2008, for more information). Most of the software for online surveys (see Kaczmirek, 2007, for an overview on software) has been developed for purposes such as market research and therefore it is not surprising that many needs in assessment contexts are not accounted for. Also, potentially useful software cannot be adapted because of its proprietary status. Hence, it was decided to develop the software testMaker to be able to administer Self-Assessments in particular as well as surveys in general.

Examples of ensuring quality of software in practice

According to the quality criteria mentioned above, some examples of means to ensure quality in the design and development of testMaker are described below. Words in brackets will link the examples to the sub-characteristics of ISO/IEC 9126.

First of all, in order to ensure *Functionality*, a dutybook had been written to define the needs of the assessment as mentioned above. This document has served as a basis for development, but also for evaluation, because it allowed a comparison between user objectives and functions of the software

(Suitability). Furthermore, an external programmer was occupied to hack into the software to find out if it was possible for unauthorized people to gain access to data which is certainly especially a problem in the assessment context (Security).

A special topic within *Reliability* is the handling of errors (Fault tolerance) as it can never be guaranteed that the software is completely free of them, but they should be removed as soon as possible after having been detected. If an error occurs in testMaker, the user will be informed about this incidence and the fact that an e-mail has been sent to the administrator. The user can then just go on in the usage of the software. If the error occurs again in the system within a defined number of days, no new e-mail is written to prevent misuse of this feature, e.g. by somebody causing errors on purpose.

An aspect that is often neglected in software is the graphical user interface although it is one of the key components of *Usability*. In the design of testMaker, a lot of attention was paid to creating a user interface that is easy to learn (Learnability), e.g. by using a tree structure known from other applications to represent the elements of a test. Another example is that no programming knowledge is necessary to operate the system, because all inputs can be made via understandable buttons, for instance in the text editor for the formatting of items and instruction pages, or placeholders for the creation of dynamic feedback pages. Since assessment projects teams mostly consist of several members, this enables each member to quickly learn how to use the system and to check the correct implementation of the assessment. Documentation of testMaker consists of an overview of all features, a short introduction into the system as well as context-specific help pages for test authors on the one hand, and an overview over the structure and annotation of the source code for programmers on the other hand (Understandability).

As to *Efficiency*, simulations resembling the participation of users in an assessment were run to analyze the response and processing times of the software (Time behavior) and the amount of load on the server (Resource utilization). Thereby, it was assured that testMaker could be used for assessments with a large number of participations at the same time.

Using the software over a long period of time like in long-term assessments raises questions about *Maintainability*. Additional requirements based on experiences and further developments made it necessary to include new features in the software. This always bears the risk of building in new errors which might not show up in software tests or if they do show up cannot easily be traced back to their cause. Therefore, an error report system was developed for testMaker that creates a detailed report about where, when and under which circumstances the error occurred and which therefore contains useful information for programmers in order to repair it (Analyzability). Also, newer versions of testMaker are always compatible to older data (Stability), because potential changes in the structure of data are adjusted by the software.

In the context of newer versions, the topic of the installation of the software comes to mind (Installability) as part of the criterion *Portability*. The installation of (newer versions of) testMaker requires access to a database and a webserver to which the software needs to be copied, but except for that, the installation can be done by any user via a graphical user interface. Keeping this process very simple prevents mistakes in the installation that could affect the operation of the software.

Added value of Open Source software

All the remarks about quality criteria mentioned above are true for any kind of software used in the assessment, there is no need to define new quality criteria for Open Source software. What is different, though, about Open Source software is the development process which influences how quality is considered and will be discussed in the following.

Three major characteristics of the Open Source definition are (1) that the software can be used, copied and disseminated as much as one likes to, (2) that the software or the source code respectively are available in a readable and understandable form and (3) that the software may be changed and disseminated in the changed way. These freedoms have many positive consequences for the use in practice. For example, the software can be adjusted to the own needs independent of a third party like the owner. Also, transparency is increased because the viewable source code enables

everyone to conduct own verifications. Additionally, Open Source software includes economic benefits on a microeconomic level (no expenses for licenses) and on a macroeconomic level (no redundant programming).
These characteristics of Open Source software and their consequences have direct and indirect effects on the quality of the software.

As illustrated in Figure 1, in general, it can be stated that higher quantity and quality of development improve the quality of the software which increases the quantity of use. This again raises the quantity of software tests influencing the quantity and quality of development (Figure 1).

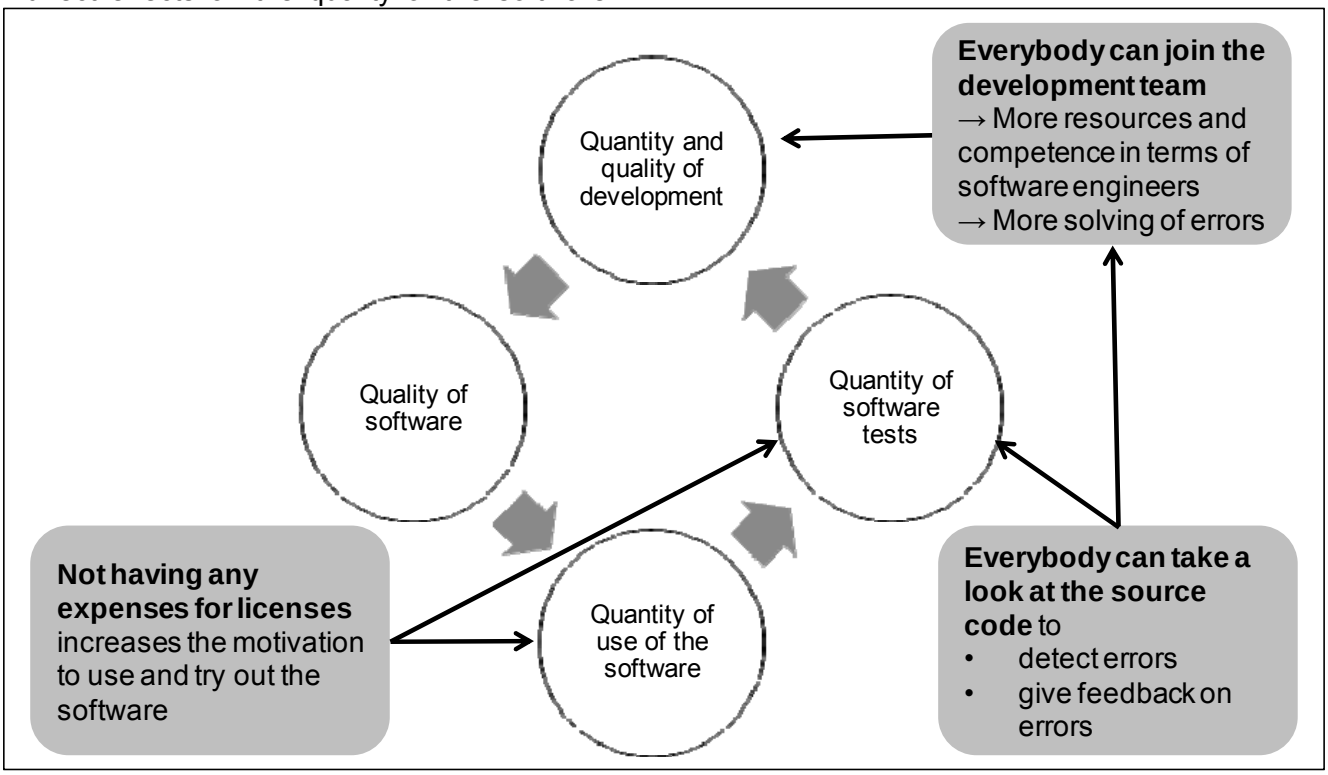


Figure 1 How Open Source characteristics influence software

Since Open Source software does not require expenses for licenses, the quantity of use and of tests are increased. Because everybody has free access to the source code, software tests are not just done by the original software development team, but any interested programmer. In this case, the probability of more errors being detected and reported is growing. Hereby, the development team increases in resources and competence to solve errors.

Unfortunately, many times Open Source software projects fail to work like that. They never gain attractiveness to others or lose it again so the software stays fragmentary or at a very trivial level. So a few points shall be mentioned that characterize successful Open Source software projects. First of all, they mostly consist of a development team of several programmers instead of a single person. This serves reciprocal control and prevents end or failure of software development if one developer leaves. Another trait is hierarchic organization of the development team. That means that one or more developers are in the function of moderators and decide about the basic goals and direction of the development and keep the software development going on. They also ensure control of quality by making tests of the software which adds up to that the central developing team needs to have a lot of knowledge and software engineering competence. Documentation is also a big issue, because it facilitates new programmers to join the team. If software is not just from programmers for other programmers, usability of the software is important to make the software attractive for other users not so familiar with programming. Finally, a communication and work platform needs to exist to allow both users and programmers to exchange files and information about the software. If planning on using Open Source software for assessments, these points can help to recognize an active, successful Open Source software project. Only such a project promises to ensure quality in software.

In technical contexts, quality is the degree to which an entity, e.g. the software product, fulfills requirements and needs. It can only be optimized, but never be guaranteed. In this sense, making sure that the software is of high-quality is a necessary precondition for a high-quality

assessment. In summary, one of the main goals in assessments is to reduce error variance in the data in order to gain a reliable measure of the true variance – improving quality of the software used in computer-based assessment is one way to do this.

References

DIN Deutsches Institut für Normung e.V. (2002). DIN 33430. Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen. Berlin: Beuth-Verlag.

International Organization for Standardization (2001-2004). ISO/IEC 9126 Software sengineering.

International Test Commission (2005). International Guidelines on Computer-Based and Internet Delivered Testing. URL: http://www.intestcom.org/Downloads/ITC%20Guidelines%20on%20Computer%20-%20version%202005%20approved.pdf [Status: 25.02.2007].

Kaczmirek, L. (2007). Online-Befragungen: Software. URL: http://www.gesis.org/Methodenberatung/Datenerhebung/Online/software.htm [Status: 26.11.2007].

Milbradt, A. & Putz, D. (2008). Technische Herausforderungen bei Self-Assessments. In H. Schuler & B. Hell (Hrsg.), Studienberatung und Studierendenauswahl (pp. 102-109). Göttingen: Hogrefe.

Milbradt, A., Zimmerhofer, A. & Hornke, L. F. (2007). *testMaker* - a software for web-based assessments [Computer software]. Aachen: RWTH Aachen University, Department of Industrial and Organizational Psychology.

Weber, V. (2003). Neukonstruktion und erste Erprobung eines webbasierten Self-Assessments zur Feststellung der Studieneignung für die Fächer Elektrotechnik, Technische Informatik sowie Informatik an der RWTH Aachen. RWTH Aachen, Aachen.

Westhoff, K., Hellfritsch, L., Hornke, L., Kubinger, K., Lang, F., Moosbrugger, H., Püschel, A. & Reimann, G. (2004). Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430. Lengerich: Pabst Science Publishers.

Zimmerhofer, A. (2003). Neukonstruktion und erste Erprobung eines webbasierten Self-Assessments zur Feststellung der Studieneignung für die Fächer Elektrotechnik, Technische Informatik sowie Informatik an der RWTH Aachen. RWTH Aachen, Aachen.

Zimmerhofer, A., Heukamp, V. M. & Hornke, L. F. (2006). Ein Schritt zur fundierten Studienfachwahl - Webbasierte Self-Assessments in der Praxis. Report Psychologie, 31 (2), 62-72.

The author:

Dipl.-Psych. Annika Milbradt
RWTH Aachen University
Department of Industrial and Organizational Psychology
Jaegerstr. 17/19
D-52066 Aachen

E-Mail: Annika.Milbradt@psych.rwth-aachen.de

WWW:
http://www.assess.rwth-aachen.de
http://www.global-assess.rwth-aachen.de

The author is a research assistant at the Department of Industrial and Organizational Psychology at RWTH Aachen University. Since 2005 she is engaged in several projects focusing on development, implementation and evaluation of web-based Self-Assessments for prospective university students. Besides, she is responsible for design and development of an Open Source software for web-based assessments, testMaker. Other research interests include online surveys in general and computer-based diagnostics of planning ability.

Quality features of *TCExam*, An Open Source Computer-Based Assessment Software

Nicola Asuni
Tecnick.com s.r.l.

Abstract
General advantages of Computer-Based Assessment (CBA) systems over traditional Pen-and-Paper Testing (PPT) have been demonstrated in several comparative works. Scientific literature generally tends to be very poor in identifying a set of criteria that may be useful to select the most appropriate CBA tool for a specific task and a lot of work is still necessary to analyse all the issues when choosing and implementing a CBA tool, even if a relevant effort has been made in this field with the ISO9126 standard for "Information Technology – Software Quality Characteristics and Sub-characteristics". In this paper, I propose to take into consideration the specific quality features of TCExam, not included in ISO9126 but extremely relevant for CBA design. TCExam is a simple, free, Web-based and Open-Source CBA system that enables educators and trainers to author, schedule, deliver, and report on surveys, quizzes, tests and exams. The paper discusses some quality features of the TCExam without entering in-depth software functionality details.

Introduction

Computer-based assessment (CBA), also known as *Computer-based testing* (CBT) or *e-exam,* has been available in various forms for more than four decades. In the past dozen years, CBA has grown from its initial focus on certification testing for the IT industry, to a widely accepted delivery model serving elements of virtually every market that was once dominated by Paper-and-Pencil Testing (PPT). Today, nearly one million tests per month are delivered in high-stakes, technology-enabled testing centres in all over the world (Tomson Prometric, 2005).

Several comparative works in scientific literature confirm the general advantages of CBA systems over traditional PPT (Vrabel, 2004). These general advantages include: increased delivery, administration and scoring efficiency; reduced costs for many elements of the testing lifecycle; improved test security resulting from electronic transmission and encryption; consistency and reliability; faster and more controlled test revision process with shorter response time; faster decision-making as the result of immediate scoring and reporting; unbiased test administration and scoring; fewer response entry and recognition errors; fewer comprehension errors caused by the testing process; improved translation and localization with universal availability of content; new advanced and flexible item types; increased candidate acceptance and satisfaction; an evolutionary step toward future testing methodologies.

While CBA is now an accepted testing solution there are still many factors that must be considered when choosing and implementing a assessment solution. The scientific literature is very poor in respect of identifying a set of criteria that may be useful to an educational team wishing to select the most appropriate tool for their assessment needs. Relevant help is provided in this direction by a number of research studies in the field of Software Engineering providing general criteria that may be used to evaluate software systems (Valenti et al, 2002). Furthermore, progress has been made, in this field by the International Standard Organization that in 1991 defined the ISO9126 standard for "Information Technology – Software Quality Characteristics and Sub-characteristics" (ISO, 1991). The ISO9126 standard is a quality model for product assessment that identifies six quality characteristics: functionality, usability, reliability, efficiency, portability and maintainability. Each of these characteristics is further decomposed into a set of sub characteristics. Thus, functionality is characterised by the categories suitability, accuracy, interoperability, compliance and security.

Nowadays several CBA tools are available on the market, but unfortunately most of them are proprietary, closed, centralized, complex, expensive and do not fully cover the aforementioned ISO9126 quality model. This is

why the author decided to start the *TCExam* project, a simple, free, web-based and Open-Source CBA system that enables educators and trainers to author, schedule, deliver, and report on surveys, quizzes, tests and exams. *TCExam* project was started in 2004 and now it is translated in several languages and freely used all over the world by universities, schools, private companies and independent teachers.

In this paper I propose to take into consideration specific quality features of the *TCExam* software, not included in ISO9126 but extremely relevant for CBA design. After a brief introduction to the tool, the proposed quality features will be described in detail and finally discussed.

TCExam general information

TCExam (http://www.tcexam.com) is a free Web-based and Open-Source Computer-Based Assessment (CBA) software application hosted on the *SourceForge.net* repository.

TCExam is divided into two main sections: public and administration. The public area contains the forms and the interfaces that will be used by users to execute the tests. In order to access this area, the users must login, inserting their username and password in the specific form. Once logged in, the users will see a page with the list of the tests to complete, and possibly the tests already done. The list of tests visualized depends on the relative time frames, the user IP address, the user's group and the condition if they have already been performed or not. The list of active tests shows, other than the test name, a list of links which can be different case by case: *info* – to display test information; *execute* – to start the test; *continue* – to continue previously interrupted test; *results* – to display test results (*TCExam* automatically grades the users' answers in real-time, considering the question difficulty and the test base score).

The test execution form contains two sections. In the first section the user may answer the selected question. The second section contains a menu to select the questions and display their status (selected, displayed, answered, difficulty). The user is freely allowed to change the answers at any time during the test. Users may leave a general comment to the test and also terminate the test at any time. It is not necessary to confirm the end of the test since it is considered to be concluded when the expiration time has been reached.

The administration area contains the forms and the interfaces to manage the whole system, including the user and database management, the generation of the tests and the results. The access to the various administration sections depends upon the user's level and group. The test-takers activity could be monitored in real time by administrators. An administrator has the privileges to stop, restart or increase the remaining time of each test. Once a test is completed, an administrator can: manually grade the TEXT answers; display, export (CSV, PDF) and print the general and detailed results; send the results to each user by email; display the test statistics. *TCExam* may also generate tests in PDF format to be printed and used in a traditional Pen and Paper Testing (PPT).

Currently *TCExam* support four question types:

- MCSA (Multiple Choice Single Answer): The test taker can only specify one correct answer (radiobutton).
- MCMA (Multiple Choice Multiple Answer): The test taker may select all answers that apply (checkbox).
- ORDER (Ordering Answers): The test taker has to select the right order of the alternative answers.
- TEXT (free-answer questions, essay questions, subjective questions, short-answer questions): Answer can be a word, phrase, sentence, paragraph or lengthy essay. Essay questions are scored manually. Short-answers are automatically graded.

Since *TCExam* is in continuous development, additional question types will probably be added in the future.

TCExam Quality Features

In addition to the aforementioned ISO9126 quality model and general CBA features, *TCExam* introduces other specific quality features that are discussed in this section.

Free and Open Source

Open Source promotes software reliability and quality by supporting independent peer review and rapid evolution of source code. *TCExam* is

a Free Libre Open Source Software (FLOSS) by adopting the GNU-GPL (General Public License). The general advantages derived by the Open Source model adoption are (Wieërs, 2008):

- Openness: All advantages of Open Source are a result of its openness. Having the code makes it easy to resolve problems (by you or someone else). You, therefore, don't have to rely on only one vendor for fixing potential problems.
- Stability: Since you can rely on anyone and since the license states that any modification shipped elsewhere should be equally open, this means that after a period of time Open Source software is more stable then most commercially distributed software.
- Adaptability: Open Source software means Open Standards, thus it is easy to adapt software to work closely with other Open Source software and even closed protocols and proprietary applications. This solves vendor lock-in situations which “ties your hands and knees” to one and only one vendor if you choose one's products.
- Quality: A wide community of users and developers does not only ensure stability, but also supplies new possibilities, making Open Source software a feature-rich solution. New features, less bugs and a broader (testing) audience (peer-review) are significant to the quality of a product.
- Innovation: Competition drives innovation and Open Source keeps competition alive. As no-one has an unfair advantage, everybody has the possibility to add value and provide services.
- Security: It is widely known that security by obscurity is not a secure practice in the long run. By opening the code and by wide adoption of Open Source software, it grows more secure.
- Zero-price: *TCExam* software is freely available and doesn't cost any additional licenses per user/year. This is probably why *TCExam* is more used on developing countries.

Community Support

The *TCExam* project is managed and distributed trough the *SourceForge.net* repository. *SourceForge.net* is currently the world's largest Open Source software development web site. *SourceForge.net* provides free hosting to Open Source software development projects with a centralized resource for managing projects, issues, communications, and code. Through the *SourceForge.net* Web site, *TCExam* users can download the latest version, read the latest news, get support, submit bugs, submit patches or request new features.

The community support is an important part of the *TCExam* development process. *TCExam* is in continuing development to reflect the real needs of the users and improve all aspects of the software quality.

Platform Independent

TCExam is a Web-based application developed on the popular LAMP platform (GNU-Linux Operative System, Apache Web server, MySQL Database Management System and PHP programming language). Part of *TCExam*’s attraction is that it can be installed on almost any server that can run PHP, including Unix, Solaris, Mac OS X and Windows systems. The database is fully documented to be easily extended or accessed by external applications. In addition, PostgreSQL can be used instead of MySQL and it is also possible to add drivers for other DBMS. No additional commercial or expensive software is required to run *TCExam*. This gives *TCExam* great installation flexibility in existing environments (i.e. a PC on a school computer room or a commercial remote Web-Server).

TCExam uses a common Three-Tier structure as in figure 1. Administration and public areas are physically separated on file system to improve security.

As a Web-based application, *TCExam* runs on a Web server and uses Web pages as the user interface. For users, all *TCExam* requires is a computer or PDA with a Web browser (i.e. Mozilla Firefox or Internet Explorer) and an Internet or Intranet connection to the *TCExam* Web server. No additional software or specific hardware is required to use *TCExam*.

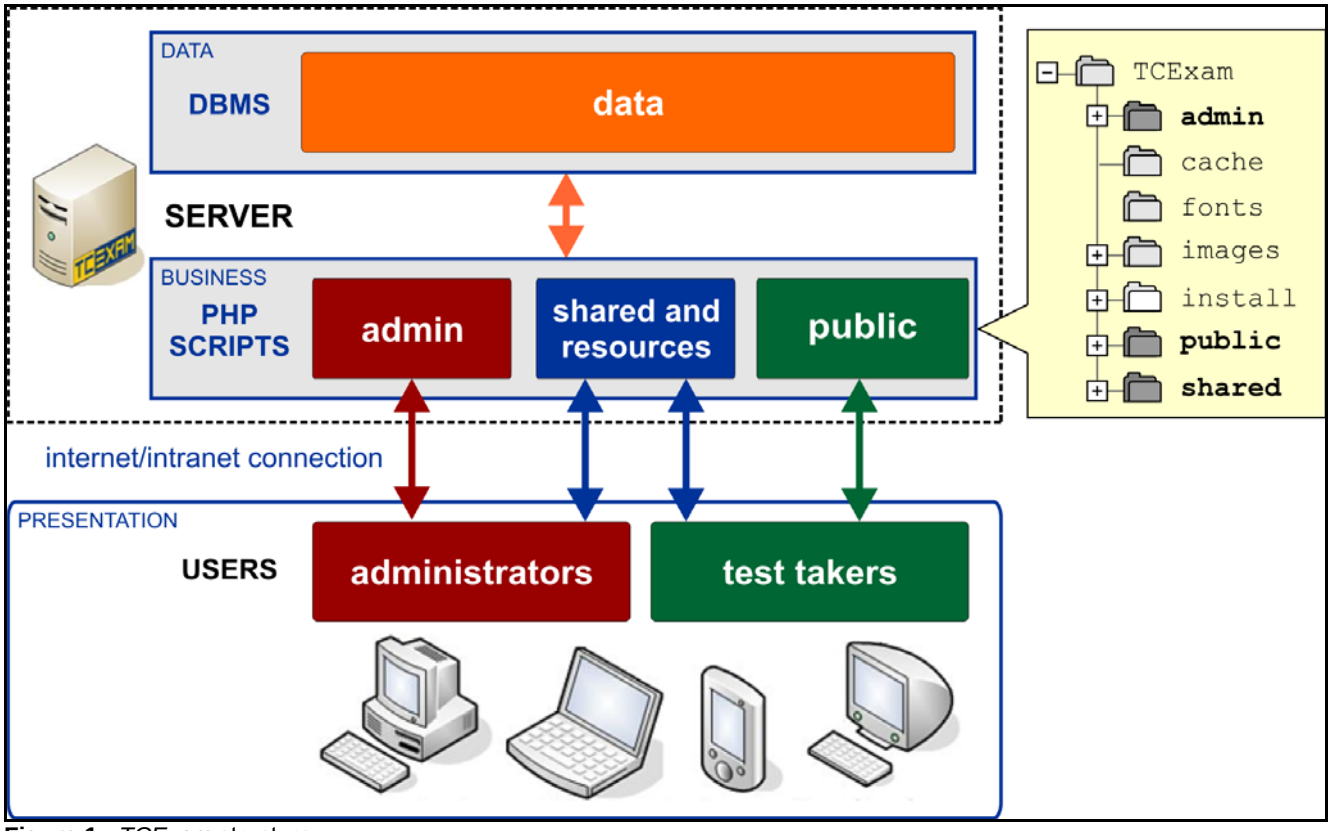


Figure 1 - *TCExam* structure.

No Expensive Hardware Requirements
The LAMP platform and the flexible technical requirements make it possible to install *TCExam* on almost any computer and even run it on shared Web servers managed by Web hosting providers. Experimental results show that a five years old PC, based on AMD Athlon XP 2400+ processor, 1GB RAM and a 100Mbps Ethernet card, may easily handle 50 tests at the same time. This feature is particularly important to bridge the gap of the digital divide with developing countries or rural areas, where modern hardware is unavailable or too expensive.

Internationalization (I18N)
TCExam is language independent by adopting the UTF-8 Unicode charset (Unicode Inc, 2005) and TMX (Translation Memory eXchange) standard (Savourel, 2004). TMX (Translation Memory eXchange) is the vendor-neutral open XML standard for the exchange of Translation Memory (TM) data created by Computer Aided Translation (CAT) and localization tools. The purpose of TMX is to allow easier exchange of translation memory data between tools and/or translation vendors with little or no loss of critical data during the process. All *TCExam* translations are included in a single XML file that could be easily edited manually or with a dedicated CAT tool. In this way everyone may download *TCExam* and add a new language translation without waiting the next software release.

TCExam supports Right-To-Left languages (i.e. Arabic, Hebrew, Persian) and already includes translations in several languages. The users may change the interface language at any time by using the selector at the end of each page.

Accessibility and Usability
It is essential that CBA tools be accessible in order to provide equal access and equal opportunity to people with disabilities. *TCExam* generates Web interfaces that conform to the XHTML 1.0 Strict standard (Pemberton et al, 2000) and W3C-WAI-WCAG 1.0 Accessibility (Chisholm et al, 1999) and Usability (US Department of Health and Human Services, 2005) guidelines. The graphic aspect of the user's interfaces is fully handled by CSS level 2 style sheets (Bos et al, 1998). CSS benefits

accessibility primarily by separating document structure from presentation (Jacobs et al, 1999). Style sheets were designed to allow precise control - outside of mark-up - of character spacing, text alignment, object position on the page, audio and speech output, font characteristics, etc.

Accessibility means that people with disabilities can use the *TCExam*. More specifically, means that people with disabilities can perceive, understand, navigate, and interact with the *TCExam* software. Accessibility also benefits others, including people with "temporary disabilities" such as a broken arm, and people with changing abilities due to aging. Web accessibility encompasses all disabilities that affect access to the Web, including visual, auditory, physical, speech, cognitive, and neurological disabilities. Web accessibility also benefits people without disabilities in certain situations, such as people using a slow Internet connection.

Usability measures the quality of a user's experience when interacting with the software application. In general, usability refers to how well users can learn and use a product to achieve their goals and how satisfied they are with that process. It is important to realize that usability is not a single, one-dimensional property of a user interface. Usability is a combination of factors including:

- Ease of learning - How fast can a user who has never seen the user interface before learn it sufficiently well to accomplish basic tasks?
- Efficiency of use - Once an experienced user has learned to use the system, how fast can he or she accomplish tasks?
- Memorability - If a user has used the system before, can he or she remember enough to use it effectively the next time or does the user have to start over again learning everything?
- Error frequency and severity - How often do users make errors while using the system, how serious are these errors, and how do users recover from these errors?
- Subjective satisfaction - How much does the user like using the system?

With the support of the University of Bologna, *TCExam* has been successfully tuned to be easily used by blind users.

Data Import and Export

To improve the software flexibility and compatibility with other CBA software, e-learning applications or existing databases, *TCExam* includes some tools to directly export or import users, questions or results data using various open formats: CSV (Comma Separated Values), XML (eXtensible Mark-up Language) and PDF (Portable Document Format). The detailed results in PDF format can be automatically sent by e-mail to each user. In addition, the database is fully documented in order to make it easily accessible by external applications (i.e. phpMyAdmin) to perform custom data import/export or backup procedures.

The current *TCExam* version includes RADIUS (Remote Authentication Dial In User Service) and LDAP (Lightweight Directory Access Protocol) modules, to directly access existing large database of users. Other authentication modules can be easily added to *TCExam* to meet specific needs.

Rich Content

TCExam uses a custom mark-up language to add text formatting, images, multimedia objects (audio and video) and mathematical formulas (supports LaTeX). *TCExam* includes a simple graphic interface with buttons to easily format the text or add external objects (i.e. images, audio files, videos, flash animations, etc). Generally, any object that could be rendered with a Web browser using a specific plug-in can be added to the *TCExam* questions, alternative answers or general descriptions.

The mark-up language used by *TCExam* is similar to the common BBCode (Bulletin Board Code), the lightweight mark-up language used to format posts in many message boards. The available tags are indicated by rectangular brackets surrounding a keyword, and they are parsed by the *TCExam* system before being translated into a XHTML or PDF. The *TCExam* mark-up code was devised to provide a safer, easier and more limited way of allowing users to format their content.

Using the special tag “[tex]” or TEX button it’s possible to add LaTeX code to represent mathematical formulas, tables or graphs. LaTeX is a document preparation system for high-quality typesetting. It is most often used for medium-to-large technical or scientific

documents but it can be used for almost any form of publishing. The *TCExam* LaTeX renderer converts the code to a PNG image to be displayed or printed.

Unique Test
In *TCExam*, questions are grouped into topics. *TCExam* can store an unlimited number of topics. Each topic can contain an unlimited number of questions and each question can have an unlimited number of alternative answers. A *TCExam* test can include several topics. For each topic or group of topics *TCExam* randomly extracts a specified number of questions with certain characteristics (i.e.: question type, question difficulty and number of alternative answers to be displayed). If the question bank is large enough, *TCExam* may generate unique test for each user by randomly selecting and ordering questions and alternative answers. This drastically reduces or eliminates the risk of copying between users.

Final remarks

The interest in Computer-Based Assessment systems has increased in recent years, and this raise the problem of identifying a set of quality criteria that may be useful to an educational team wishing to select the most appropriate tool for their assessment needs. In this paper I have proposed to take into consideration the specific quality features of specific CBA software called *TCExam*, in addition to the ISO9126 software quality model. The proposed quality features not only improves the quality of CBA software but also positively influence its diffusion and developing model.

References

Asuni, N. (2004), PHP Localization with TMX standard, PHP Solutions Nr 3/2006.

Bos, B., Lie, H.W., Lilley, C., Jacobs, I. (1998), Cascading Style Sheets, level 2 - CSS2 Specification, W3C, http://www.w3.org/TR/REC-CSS2

Chisholm, W., Vanderheiden, G., Jacobs, I. (1999), Web Content Accessibility Guidelines 1.0, W3C http://www.w3.org/TR/WCAG10/

ISO (1991). Information Technology – Software quality characteristics and sub-characteristics. ISO/IEC 9126-1.

Jacobs, I., Brewer, J. (1999) Accessibility Features of CSS, W3C, http://www.w3.org/TR/CSS-access

Pemberton, S., et al (2000), XHTML™ 1.0: The Extensible HyperText Mark-up Language, W3C, http://www.w3.org/TR/2000/REC-xhtml1-20000126/

Savourel, Y. (2004), TMX 1.4b Specification, The Localisation Industry Standards Association (LISA), http://www.lisa.org/tmx/tmx.htm

Tomson Prometric (2005), The Benefits and Best Practices of Computer-based Testing, Tomson Prometric, ThomsonPrometricBestPractices.pdf

Unicode Inc (2005), What is Unicode?, http://www.unicode.org/standard/WhatIsUnicode.html

US Department of Health and Human Services (2005), Usability Basics, Usability.gov, http://www.usability.gov/basics/index.html

Valenti, S., Cucchiarelli, A., Panti, M. (2002), Computer Based Assessment Systems Evaluation via the ISO9126 Quality Model, Journal of Information Technology Education Volume 1 No. 3.

Vrabel, M. (2004), Computerized versus paper-and-pencil testing methods for a nursing certification examination: a review of the literature, Comput Inform Nurs. 2004 Mar-Apr;22(2):94-8; quiz 99-100. Review.

Wieërs, D. (2008), Open Source advantages, Linux.be, http://linux.iguana.be/open-source/

The author:
Nicola Asuni
Tecnick.com s.r.l.
Via Della Pace, 11
09044 Quartucciu (CA) – Italy
E-Mail: nicola.asuni@tecnick.com
WWW: http://www.tcexam.com

Nicola Asuni is the founder and president of Tecnick.com s.r.l., a provider of IT services and Open Source Software used by millions of people around the world. He holds a Master of Science degree in Computer Science and Technologies from the University of Cagliari, Italy. He has been a freelance programmer since 1993 and has actively contributed to several popular Open-Source projects. He has been also involved in research activities on Computer-Based Assessment, Web applications and interfaces, data acquisition systems and computer imaging.

An Open Source and Large-Scale Computer Based Assessment Platform : A real Winner

Matthieu Farcot & Thibaud Latour
CRP Henri Tudor

The domain of assessment is inherently wide and highly multi-form. This variability is largely reflected in the diversity of existing software tools that support it. Taking into account this dramatic variability of contexts, it is clear that the approach adopted so far for the development of computer-assisted tests, i.e., on a test-by-test basis or focusing on a unique family of competencies, is no longer viable. Only a platform approach where the focus is put on the management of the whole assessment process enables covering consistently the entire domain. In addition, this platform should rely on advanced technologies ensuring adaptability, extensibility, and versatility at user-level. Since the organizational complexity of stakeholders in the assessment process may be particularly large, collaborative and distributed aspects should not be underestimated both in the functional space of the software and in the development process itself.

TAO is a dedicated large-scale computer based assessment (CBA) platform developed jointly by the Public Research Centre Henri Tudor (Centre for IT Innovation – CITI) and the University of Luxembourg (Educational Measurement and Applied Cognitive Science – EMACS). This trans-disciplinary approach to the design and development of TAO resulted in the creation of a large scale CBA generic platform independent of any specific context of use. The collaborative framework relied on a strong iterative approach for its development, as computer sciences were adequately complemented with psychometric expertise. This development process led to the creation of a working prototype, and incited the German Deutsches Institut für Internationale Pädagogische Forschung (DIPF) to join the project.

The current TAO platform consists in a series of interconnected modules dedicated to Subject, Group, Item, Test, Planning, and Result management in a peer-to-peer (P2P) network. Each module is a specialisation of a more generic kernel application called Generis **d**eveloped in the framework of the project. The specialisation consists in defining the domain of specialisation by adding a model to the kernel, several plug-ins providing domain-dependent functionalities relying on the model, possibly some external applications, and a specific (optional) user graphical interface.

Offering versatility and generality with respect to contexts of use requires a more abstract design of the platform and well-defined extension and specialisation mechanisms. One of the main design challenges is to enable the user to create their own models of the various CBA domains while ensuring rich exploitation of the meta-data produced in reference of these models. Semantic Web technologies are particularly suited in this context, and have been used to manage the whole CBA process and the user-made characterisation of all the assessment process resources. This layer is entirely controlled by the user and includes distributed ontology management tools (creation, modification, instantiation, sharing of models, reference to distant models, query services on models and meta-data, communication protocol, …). Each modules of this layer is built as a specialisation of a generic kernel, called Generis, providing modelling and data sharing related services (Plichart et al. 2004).

Such architecture, together with semantic query services available on each module enables very advanced and useful functionalities for result analysis. Indeed, rich correlations can be made between test execution results (such as scores and behavioural observations) and any element of the meta-data defined by the user on all resources of the CBA process and collected in the entire module network.

The very core of TAO is initially rooted in an academic framework. These knowledgeable actors are traditionally prone to adopt an “Open Science” approach to research and development. Such an approach values knowledge exchange and peer review

processes as the best ways to achieve excellence in asset creation.

This naturally led to the decision to use an Open Source license for the platform. As it shall be demonstrated below, this decision is not only rational in view of the habits of the team members working on the project. This licensing choice is also the most rational decision the management team could take for the creation of this software platform considering both the technical and economical environments of large scale CBA software.

Economics of Open Source in the light of large scale CBA

An analysis of the organizational scheme used in Open Source software has been clearly described by E. von Hippel and G. von Krogh [2003]. These authors oppose the two prevalent models generally used for intellectual asset development, namely "collective action" versus "private investment" models. Whereas the former is a means of explaining the specificities of the "Open Science" approach referred to above, the latter is a more conventional incentive to create with a view to obtaining exclusive ownership rights over an invention or a creation. The authors qualify this distinction by proposing a specific type of model to characterize Open Source projects, i.e. the "private collective" model, containing elements from both of these models. A private incentive, based on the prospect of economical returns is thus completed by a common creation shared by all potential users and suppliers. This scheme benefits all concerned by motivating competitors to cooperate in what is known as a "coopetitive" framework so as to develop the best product at the lowest cost for each player.

As stated by Tirole and Lerner [2002]: "When can it be advantageous for a commercial company to release proprietary code under Open Source license? The first condition is (…) that the company expects to boost its profits on a complementary segment. A second is that the increase in profit in the proprietary complementary segment offsets any profit that would have been made in the primary segment, has it not been converted to Open Source. Thus, the temptation to go Open Source is particularly strong when the company is too small to compete commercially in the primary segment or when it is lagging behing (…)."

TAO as a business opportunity aggregator

By a simple analogy, this situation can be compared to a large suburban shopping centre offering free parking: consumers will be drawn by the provision of a free commodity to take advantage of the opportunities available to them. The more opportunities offered, the greater the value of the asset. Over and beyond this simile, software and related services are by definition intangible goods, therefore not subject to physical constraints such as rivalry in use. The same software can be widely disseminated and used without exposing consumers to any potential losses. On the contrary, the wider the diffusion the more value the software acquires through network effects.

Ongoing Open Source software projects follow strong and specific dynamic trajectories. The collateral effects of the freedom to use the software make it possible for any actor to initiate a "fork": that is, derivative work based on the original but no longer controlled by the initial creators. Another feature of this specific dynamic trajectory is linked to community management. Thereby, those concerned can freely chose to either join or leave the virtual development team.

Created intellectual property is not easily controlled as forks can occur and knowledge can be taken away as external core team members become more or less involved. This maximizes opportunities for "spill-over" effects, uncontrolled diffusion of the software and related knowledge. However, whereas such side effects are perceived negatively should the product be solely valorised though a proprietary and closed strategy, such is not the case with Open Source licensing. It should be recalled that this occurs in the framework of a global dynamic environment, with movable frontiers and optimized knowledge exchanges to the advantage –in the end– of all concerned.

TAO as a goodwill magnet

Assessment and technology experts acknowledge the fact that underlying processes and requirements for a successful Technology Based Assessment are highly multiform and reflect a tremendous diversity of needs and practices. In particular, this involves a whole range of specialists from researchers in psychometrics, educational measurement, or experimental psychology to large-scale assessment and monitoring professionals as well as educators and human resource managers.

This heterogeneity is a yet an additional asset enhancing the value of an Open Source Large scale CBA platform. Relying on open standards, and adequately generic, TAO offers a common core base onto which flexible third party business models can be connected. This is achieved by means of an agility-based model, where the common platform is commoditized by contributions from all concerned (bug fixes, feature enhancements, …). They also have a clear incentive to develop their individual software CBA solutions using the best possible common platform. Such adaptability allows TAO to create and exploit new markets faster than competing, non Open Source products - using all available goodwill. “Agility” constitutes such an innovative and expanded business ecosystem.

Economic issues are of greatest importance and depend largely on architectural and organisational choices when implementing large-scale CBA. For a long time already, the CBA processes of test and item production, delivery, and analysis have been described as strictly replicating the paper and pencil practices – but often at a higher cost. These negative effects are worsened by low reusability and limited adaptability of the by-products.

TAO as a disruptive advantage for end users

This global scheme constitutes the disruptive advantage of TAO within the large-scale CBA market. Open Source solutions dedicated to CBA exist, but they usually offer a low level of interactivity and usability for test creators. On the contrary, Open Source large scale assessment solutions are very scarce. This market is mainly dominated by a few proprietary solutions that rely on strong user lock-ins which hinder the diffusion of such products, owing mostly to their costs and deliberate lack of interoperability. The “platform fits all” strategy followed by TAO is therefore a unique solution on the market.

It reduces time-to-delivery of tests by enabling a reuse of not only the content, but also of any software components already produced and shared by the community.

It induces mid-and long-term cost reductions by minimizing the development needs for new specific components. Therefore investments are limited to a one off call. This is mainly possible because of the shift in term of market power from the hands of the solution providers to those of the users who can then share their developments with each other. The open licensing scheme followed is an essential enabler of such a positive dynamic practice.

Which Open Source license for TAO?

As stated by Stallman [2002, p.20] : “Copyleft uses copyright law, but flips it over to serve the opposite of its usual purpose : instead of a means of privatizing software, it becomes a means of keeping software free”.

One of the best available methods for Open Source licensing to achieve the maximum benefit of a Copyleft policy is the GNU General Public License v2 (GPL v2). This license authorises anyone to run, copy, modify, and distribute the modified versions of the program under license as long as the modified program derived from the original one remains within the terms of the same license. This “viral” clause restrains any possibility of private appropriation in case of diffusion of the modified code: any borderline changes of the code automatically place the new code within the scope of the existing license.

Such a specific use of copyrights through licensing allows the creation of a real common good, composed of some original software elements and all the other derived ones.

Conclusion

TAO can be used for multiple purposes, from a global and systemic level to a more individual one. Recognizing the variety of intended uses markets and needs, TAO minimizes end-user risks by resorting to strong dynamic exchange effects through the consolidation of quality-related actions (the more users, the better).

TAO designers strongly believe in the promises of Open Source as a means to create a sustainable technological environment and related business models. Large-scale CBA platforms do not need user lock-ins to be successful. Free choice should prevail in this very particular field of software use.

However, to succeed, TAO needs to be a mature enough solution to foster trust both from end users and solution providers. For this reason, TAO is still a project subject to restricted access that will be released under the GNU General Public License version 2 to the public. But before hand, internal project dedicated professionals will finalize an ongoing process to verify that code quality and other essential features are effective enough to answer the needs and meet the expectations of all.

References

Plichart P., Jadoul R., Vandenabeele L., Latour Th. (2004). TAO, A Collective Distributed Computer-Based Assessment Framework Built on Semantic Web Standards. AISTA2004, November 15-18, 2004. Luxembourg, IEEE Computer Society Digital Library.

Stallman, R. M. (2002). Free Software, Free Society. GNU Press.

Von Hippel, E., Von Krogh, G. (2003). Open Source Software and the Private-Collective Innovation Model. Organization Science, vol 14(2):209-223

Tirole, J., Lerner, J. (2002), Some Simple Economics of Open Souce, Journal of Industrial Economics, Vol. 50, No. 2, pp. 197-234

Zimmerman, J.B., Julien, N. (2006), New Approaches to IP: From Open Source to Knowledge Based Activities, DIME working Paper

The authors:
Matthieu Farcot, Thibaud Latour
CRP Henri Tudor
29, Avenue John F. Kennedy
L 1855 - Luxembourg

E-mail: matthieu.farcot@tudor.lu
Thibaud.latour@tudor.lu

WWW: www.tudor.lu

Thibaud Latour is the Head of the "Reference Systems for Certification and Modelling" Unit at the Centre for IT Innovation department of the Public Research Centre Henir Tudor. He is also Program Manager of a project portfolio dedicated to Technology-Based Assessment of skills and competencies. He obtained his M.Sc. in Chemistry in 1993 from the Computer Chemical-Physics Group of the Facultés Universitaires Notre-Dame de la Paix (FUNDP) in Namur (Belgium), working on conceptual imagery applied to supramolecular systems and in collaboration with the Queen's University at Kingston (Ontario, Canada). From 1993 to 2000, he participated in several projects exploring Artificial Neural Network (ANN) and Genetic Algorithm (GA) techniques and developing ad hoc simulation methods for solving complex problems. During the same period, he supervised a number of M.Sc. thesis work in Computational Chemistry. In 2000, he joined the Centre de Recherche Public Henri Tudor where he was in charge of an internal Knowledge Management project. In that context, he designed a knowledge base for collaborative elicitation and exploitation of research project knowledge. He is in now involved in several projects in the fields of Computer-Based Assessment, e-Learning, and Knowledge Management where Semantic Web, Knowledge Technologies, and Computational Intelligence are intensively applied. Thibaud Latour has also served in programme committees of several conferences such as ISWC, ESWC, and I-ESA and as workshop co-chair of the CAiSE'06 conference.

Matthieu Farcot has a PhD in economics. Currently working within the valorization of ICT innovation Unit of the CRP Henri Tudor, his field of study focus on intellectual property and legal issues related to innovation. He also works on IT licensing and business models applied to free and open source projects.

What software do we need? Identifying quality criteria for assessing language skills at a comparative level

Friedrich Scheuermann & Ângela Guimarães Pereira
European Commission - Joint Research Centre (IPSC)

Abstract
Due to the increasing role of computer-based assessment (CBA) in European comparative surveys there is a general interest to look deeper into the potential of software tools and to verify to what extent these tools are appropriate and transferable between subject areas. There is a variety of aspects to be considered in order to check the usefulness of software in a specific context. Research literature, specifications, standards and guidelines help to identify the relevant indicators for verifying the quality of a specific tool which in many cases has been described as "…hard to define, impossible to measure, easy to recognise." (Kitchenham, 1989). Since various research areas from different scientific disciplines are concerned, it is not surprising that most of them focus on specific aspects and do not provide a complete picture of what is needed for ensuring (or even improving) the quality of computer-based skills assessment from a user perspective. This paper presents a review of software tools for skills assessment using as a context, the languages assessment context at European level, discussing requirements. We will argue that more connection between users and developers is needed. This will certainly enhance the "external" quality of these tools, as far as their fitness for purpose is concerned.

For a number of years the European debate about software tools for skills assessment activities has largely developed. The relevance of discussing quality criteria is given by new European surveys intended to be carried out in the following years. In general, it is assumed that electronic testing could improve the effectiveness, i.e. improve identification of skills, and efficiency, by reducing costs (financial efforts, human resources etc.) but there exists still little experience on both how to carry out such test at European level, and how to define the potential role of Information and Communication technologies (ICT) for carrying out such a survey.

In general, the potential of computer-based assessment specifically and software tools features in general are largely stimulated by actual needs of potential users, the state of technological and methodological innovation and finally the given constraint in terms of available resources. Making the right decision about needed software tools for assessing skills at a European level requires substantive reflection about the overall needs and requirements for the specific assessment context. Existing quality criteria defined for this purpose are available from various sources, such as existing research literature, standards, and guidelines etc. which describe a set of possible indicators to be applied. Some quality related issues are discussed in several articles of this report (i.e. by Bartram, Milbradt, Asuni, this volume). Here, we will try to relate the software quality discussion to the specific subject area of language learning.

Background

In 2006 the European Parliament and the Council of Europe (2006) recommended that a survey on language skills as key competences for lifelong learning should be carried out. The data should feed a common reference tool to observe and promote progress in terms of the achievement of goals formulated in the "Lisbon strategy" in March 2000 (revised in 2006, see http://ec.europa.eu/
growthandjobs/) together with its follow-up declarations in selected fields (European Parliament and Council of Europe, 2006).

More specifically, the European Council conclusions (2006) on the *European Indicator of Language Competence* asks for "measures" for objective testing of skills for first and second foreign languages based on the Common European Framework of Reference for Languages (CEFR). The Council conclusions suggested assessing competence in the 4 receptive and productive skills, but "for practical reasons" to focus on the following areas in the first round: listening comprehension, reading comprehension and writing; the testing of speaking skills being left for a later stage.

As a consequence, an assessment instrument is needed for monitoring the actual state of various types of language skills in European countries (across various time intervals) and to observe improvements made. An important issue to look at, is to what extent the use of ICT can support this assessment exercise and the complete process of regular assessments in this domain.

The definition of requirements for CBA and software tools is very much encouraged by progress made in research on educational measurement and ICT development. Higher performance of hardware infrastructure, multimedia tools, as well as extended internet possibilities increases the potential of CBA. At the same time assessment methodologies advance rapidly and it is stimulating to reflect about improving the assessment quality by benefiting from new item types, computer-adaptive testing (CAT) and other upcoming methodologies.

However, the amount of human effort and costs are directly related to task design, needing to be carefully thought about and to be related to expected gains for language skills assessment. In financial terms required budgets and country contributions for carrying out the survey have to be low as more surveys have to be delivered both at country and European level in general.

In the given context of the language survey the questions to which appropriate answers are requested are as follows:

- What software is available for carrying out computer-based tests?
- To what extent can CBA software support and improve the process and overall quality of assessment?
- To what extent can software solutions provide cost-effective alternatives to traditional methods?

Review of software applications

Taking a closer look at the market as far as skills assessment tools are concerned, many different types of software applications can be identified. Some of them are covering the complete assessment process (of item/test development, delivery, analysis and reporting), however, the majority focuses on selected areas. Some of those tools represent stand-alone software products; others are based on assessment provided via the Internet.

Hence, we can identify a large number of electronic tools and services on the market supporting various kinds of assessment activities. Such tools are offered either as

- specific **modules of content/learning management systems (CMS/LMS)** that enable the management of (usually multiple-choice) items together with the administration and internet-based delivery of tests (e.g. Moodle, http://www.moodle.org), or
- **Authoring software tools** (e.g. Hot Potatoes, http://hotpot.uvic.ca),
- Software **dedicated to data collection and presentation** (e.g. OpenSurveyPilot, http://www.opensurveypilot.org/)
- **Administration software tools e.g. for documentation**, including pupil assessment administration) (e.g. Gradebook 2.0, http://www.winsite.com/bin/Info?2500000035898) or
- **Software for statistical computing and predictive analytics** (e.g. R, http://www.r-project.org/) or
- **Assessment management systems** with specific focus given to the support of summative of formative assessment of skills and learning (e.g. Questionmark™ Perception™, http://www.questionmark.com/us/perception/index.aspx).

There is not a shared definition for the categories described earlier; as far as assessment management software is concerned different terms are applied, such as *assessment software, assessment platform, assessment software system, testing software etc.* Some of these assessment tools, such as TAO (http://www.tao.lu) can be considered as **integrated software environments** due to their openness to integrate other independent software tools as system components. In TAO, all features and functionalities needed to cover the whole assessment process are available in the software environment and/or can be adapted to a survey's specific contextual requirements (see also Farcot & Latour, this volume).

Noteworthy, there is a large number of **assessment services** (e.g. Pan Testing) available via the Internet, covering a wide range of (tailor-made or standard) activities

proposed depending on specific needs. Such customer-oriented services are usually offered by commercial enterprises. In most cases underlying software tools applied are only accessible via adapted interfaces.

So far, based on literature review and internet search, more than 700 products and services (including 435 software tools, 210 commercial/fee-based Internet assessment services provided by companies, non-profit or public organisations) were identified which then have been explored and classified according to the categories defined earlier. Specific attention was given to Open Source software developments. Analysis could be broken down according to the categories mentioned above and focussed on 135 assessment packages, out of which 76 were offered on an open-source basis.

One of the reasons to go Open Source software (OSS) in these types of platforms is to try to boost through a community of users further developments. However, availability of OSS platforms for test delivery is quite limited. During the software review, a great deal of what is presented under this branding is not corresponding to that what is commonly understood as “Open Source” in terms of the availability of open source code (see for instance the OSI, http://www.opensource.org/). In many cases this software is declared as “work in progress” to be published at a later stage or, as in most cases, out of date and no longer accessible. Remarkably, 35 open source products generally identified were not accessible or out-of-date.

Finally, a limited number of open source software environments could be identified which were considered to be relevant for CBA purposes. Examples of such platforms are described in this report, such as TAO (see Farcot & Latour), TCExam (see Asuni,) and TestMaker (see Milbradt).

It is therefore important, to take a close look at that what is currently available in the market. In order to do so a protocol is needed, which supports the identification of quality criteria of software products, especially those offered on the basis of open source licence.

Quality criteria and contextual parameters

In the remaining of this chapter we will illustrate the quality issues that require attention when carrying out a quality assessment of software tools deployed in CBA; where possible, the language survey context, in particular at the European level, will be used to illustrate the quality framework suggested here.

As stated earlier, in CBA the assessment context determines the framework of quality issues to be addressed. A variety of indicators and success factors are described in the literature such as in research reports (e.g. McKenna & Bull, 2000), checklists and guidelines (e.g. Pass-it, 2005). Typically, these documents refer to specific assessment contexts.

The following contextual issues seem to be the most relevant aspects to be taken into account:

- Assessment rationale: purpose, objectives, functions, subject area etc.
- Intended measurement: subjective, objective etc.
- Assessment design: stakes, general approach, instruments, phases and timing etc.
- User groups: test publishers, developers, test takers, administrators
- User profiles: languages, special needs etc.
- Resources: IT infrastructures, networking resources available, test location etc.

As far as European surveys are concerned some of these issues require an analysis at a country level due to the heterogeneity of contextual parameters.

Box 1: Language Survey Context…

For a pan-European survey on languages, highest benefit may be achieved if all EU countries will participate in the survey. Given the limited budgets available at least as far as some EU countries are concerned, any solution proposed needs to be revised against cost-effectiveness.

For a European language survey, it will be important that very diverse ICT infrastructures are supported. Some countries might be well prepared, others less as e.g. some of the new member states which might have problems to comply with necessary hardware capacities and bandwidth if Internet-based operations are deployed.

Based on the given contextual parameters various CBA-delivery modes can be applied, such as computer-based storage and delivery (e.g. delivery via CD-ROM or other boot-devices), web-based delivery and assessment and/or paper-pencil delivery. In some cases, combinations of these delivery scenarios are offered in order to better adapt to specific contexts. If this is the case the issue of comparability of results needs to be investigated in order to avoid differences caused by the mode of delivery.

CBA software quality requirements depend upon the stage of the assessment process assisted by a software tool. Hence, if context and field of assessment are important to determine what features and what quality categories need to be addressed, the several steps of the assessment process such as test administration, design, delivery and survey reporting have to be carefully looked at.

If an ex-ante assessment of the context and process is carried out, it is possible to look at software tools with a normative perspective, and identify for each particular process and context whether existing software will fit the purpose or new software is needed.

Identification of main categories of quality will help taking decisions regarding software deployment, time and cost being important factors for the decision-making process. In many cases such decisions also take into account available licences and tools within the organisation in charge of software provision.

Box 2: Language Survey context

Software tools will need to take into account that all languages of participating countries are supported as far as the interface and test is concerned. Yet, there are several other user-related issues which need careful reflection before software-decision is made: what kind of item types are needed for measuring the different language skills to be assessed? Which features are needed in order to complement or substitute a paper-pencil test version – as far as the assessment process is concerned as well as the administrative aspects of the implementation.

In the subject field of languages, tools would have to be checked for their capabilities of assessing productive and receptive skills.

Writing skills (productive skills) such as writing of essays would not benefit a great deal if assessed via CBA since there are severe limitations of current assessment algorithms for essays. Only few existing robust products in the market (such as CriterionSM from ETS) can do this job, consisting of an algorithm for automated and immediate feedback on essay-writing performance (holistic score and annotated diagnostic feedback); at present there is no (open) source code available which could support the integration of these features in other platforms. Furthermore, according to experts of the commercial market a great deal of limitations can still be found. “Automatic computer-based marking of subjective, free-text responses still operates at basic levels of character or rule recognition. For this reason, the future of subjective testing will depend on human marking, albeit online marking or expensive researching and piloting of more advanced, essay-marking software.“ [Liam Wynne from Pearson Vue, 2007].

Speaking skills are even a more complex task as far as assessment is concerned. Here specific requirements are requested from the IT resources at the user side (e.g. speech recognition hard- and software possibilities) as well as the required bandwidth, which is extraordinary high.

Whether undertaken in an electronic mode or not, the most important challenge for assessing productive skills is, in both cases, that heavy investments needed to deliver and generate results at large-scale level. As demonstrated by PISA, the provision of open questions is rather cost-intensive. Due to the further overall benefits which can be achieved by CBA this should not provoke a general debate on whether CBA is needed or not.

Furthermore, as far as the process of CBA is concerned, Table 1 summarises relevant documents which specify quality issues for computer-based assessment:

ITC	International Guidelines on Computer-Based and Internet-Based Delivered Tests
Association of Test Publishers (ATP)	Guidelines for CBT
BS ISO/IEC 19796-1	ITLET Quality Management, Assurance and Metrics - General Approach
BS ISO/IEC 19796-2	ITLET Quality Management, Assurance and Metrics - Quality Model
BS ISO/IEC 19796-3	ITLET Quality Management Assurance and Metrics - Reference Methods and Metrics
BS ISO/IEC 19796-4:	ITLET Quality Management Assurance and Metrics - Best Practice and Implementation Guide
BS ISO/IEC 24763 (TR)	Conceptual Reference Model for Competencies and Related Objects
BS ISO/IEC 23988	Code of Practice for the use of IT in the delivery of assessments
BS7988: 2002	Code of Practice for the use of IT for the delivery of assessments

Table 1: CBA specifications, guidelines

These specifications represent different CBA-related perspectives. Some of them take a rather implementation-oriented view, others are kept very IT-focussed looking at relevant quality issues for CBA and test implementation processes in assessment.

Software quality specifications

Quality criteria for software products can be derived from indicators with direct and indirect implications for software quality. As far as direct indicators are concerned some specifications are already described in other section of this report. Due to the large number of specifications identified only a brief overview can be provided here.

The most relevant one, relating to software quality in general is ISO/IEC 9126 (see figure below) with a description of a set of categories and sub-domains to be taken into account. Some of these (such as “maintainability”) can hardly be reviewed by the end-user since they relate to internal technical processes which only can be analysed by the developer or software experts. Valenti [2006] proposes the following domain-specific aspects of ISO/IEC 9126 to take into account for reviewing quality aspects of assessment software tools: Functionality, usability and reliability which are further described more in detail by Milbradt (this volume).

More specifically, the IMS QTI (Question & Test Interoperability, see http://www.imsglobal.org/question/), the IMS LD (Learning Design, see http://www.imsglobal.org/learningdesign/) specifi-cations and the SCORM (Sharable Content Object Reference, see http://www.adlnet.gov/scorm/) model address issues relating to the context of assessment. SCORM focuses on the description of metadata for classifying contents and of relationships to other components of the software environment.

Especially IMS QTI directly addresses assessment content and structure for software development purposes. This specification is the leading item-banking standard describing a data model for tests, items (addressing issues such as variables, interactions, response processing, item body, item templates), data usage (e.g. for statistics) and meta-data. It has been widely adopted by the software community but it has also been criticised because of the limited variety of item types taken into account. Examples for “innovative” item types are described by the DIALANG project (http://www.dialang.org/intro.htm) and further explored by Scalise, K. & Gifford, B. (2006).

It would need to be carefully reflected if the standard fulfils the assessment needs as far as methodologies and innovative approaches to skills assessment are concerned. New releases of this standard have addressed this weakness and broadened the scope of item types but have not yet achieved the critical mass of *adopters*.

Table 2 presents mainly technical, specifications that respond to general software quality assurance issues pertaining to the technical processes and software features in the field of CBA.

Standards, specifications, guidelines	
ISO 9241	Ergonomic requirements for office work with visual display terminals…
ISO/CD 14756	Measurement and rating of performance of computer based software systems
ISO/IEC 9126	Software quality characteristics and metrics
ISO 9127	User documentation and cover information for consumer software packages
ISO/IEC 12119	Software packages - Quality requirements and testing
ISO/IEC 14102	Guideline for the evaluation and selection of CASE tools
ISO/IEC DIS 15026	Systems and software integrity levels
ISO/IEC DIS 14143	Functional size measurement
ISO/IEC 14598	Software product evaluation
IEC 1508	Functional safety - Safety related systems (Part 3: Software requirements)
IEEE Std 1044	Classification for software anomalies
IEEE Std.830	Software requirements specifications evaluation criteria
ITSEC	Information technology security

Table 2: Software quality specifications

Proxy quality indicators

Quality of market products can be addressed using proxy indications. A variety of aspects which potentially provide some quality-related indications are indicated in Table 3:

Product-related information	
Market-share	*What is the popularity of the product on the market in relation with tools of the same product family?*
Regularity of updates	*Are updates offered and how often?*
Warranty	*Is any kind of warranty on the product given? Are there elements/aspects which are specifically included and/or excluded from warranty?*
Support	*Is support provided? What kind, to what extent?*
Licence / Pricing	*What are the costs of purchase and updates/maintenance/ support? What does a licence include?*
Reputation	*Are the positive software reviews, prices reported about the product?*

Table 3: Product related information.

Furthermore, any information relating to the software provider (developer) can become relevant for identifying the potential quality of a product see table 4.

Provider-related information	
General qualifications	*What is the main field of expertise and services provided by the company? What is the size in terms of staff and specialised professions? Etc.*
Financial stability	*How long has the company been in business? How well is the company performing in business*
Reputation	*What is the general reputation of the company? Positive/negative reviews?*
Quality Assurance	*Does the company undertake measures to ensure quality of products and services? Etc.*

Table 4: Provider related information.

Open Source Software

Open source approaches have great potential in education (e.g. see Bacon & Dillon, 2006). Specific aspects need to be taken into account if an open source product is to be quality assured. For many reasons, OSS appears to be very attractive, yet, one major concern is expressed by the lack of sufficient quality control mechanisms, and limited life-time of software products as proven by many Open Source projects. However, there are a few issues which can be considered in order to address perceived limitations. Some examples for such criteria are listed in Table 5.

Open-Source-related information	
On-going effort	*Is there an on-going effort to further develop the product*
Reputation/ freshmeat user rating	*What is the user-rating of the tool as indicated at http://freshmeat.net/*
Volume of available documentation	*Is there and how much documentation on the product is provided?*
Number of downloads	*How many times has this tool been downloaded?*
Degree of peer review	*To what extent is the software reviewed by other developers/users?*
Degree of stakeholders involvement	*To what extent are users (such as open source developers) involved in the development, maintenance and revisions of the product?*
Repository checkouts	*Are there regular updates and software releases?*
Volume of mailing lists	*Are there mailing lists provided, and how much is being discussed?*
Number of contributors	*How many users (Developpers) contribute to the improvement/continuation of the product?*
Number of releases	*How many releases have been published?*
Support community	*To what extent is support on the software tool provided*

Table 5: OSS criteria.

Concerning pricing, it is important to check what "other costs" are associated with the use of the open source product, e.g. related to support, training, adaptations needed etc.

Reflections

In general, we can observe that there are no specific standards and specifications exist that can be applied to CBA in the field of languages. From a user-perspective existing resources do not provide a complete, clear and meaningful picture on quality relevant aspects.

Extensive reviewing of existing literature and available software tools, show that the diversity of issues coined as "quality" issues in CBA, originate in equally diverse testing or assessment contexts and processes. At present quality assessment operationalisation is fragmented i.e. albeit fitting the general purpose for which they were conceived, they are often difficult to be transposed across different assessment contexts. There seems to be a need for harmonisation of quality issues into a broader framework of computer based assessments. Whilst we firmly believe that there are no "one size, fits all" approaches, the work presented here attempts to seek for some harmonisation, as far as establishing a framework for quality assessment of tools deployed in computer based assessment.

The nature of such criteria relate to e.g., the fitness for purpose of assessment methodologies and supporting tools (from a psychological/ psychometrical, pedago-gical perspective), technical features and specifications, as well as to socio-economic aspects. However, few experiences are documented to provide a sound overall picture of the complete scope and process of effective and efficient computer-based test delivery.

The task ahead is a complex one, since a quality framework for CBA supporting software tools is not just about the technicalities of the tools and their operation, their algorithms, or the user interfaces; not just about the whole process which they support but all of this together and complemented by the heavy dependency on the assessment processes' context and culture.

References and further literature

Bacon, S. & Dillon, T. (2006). Opening education – The potential of open source approaches for education. Futurelab series. Bristol, UK.

Bergstrom et al. (2006). Defining Online Assessment for the Adult Learning Market. In: Online Assessment and Measurement. M. Hricko and S.L. Howell. Hershey, London, Information Science Publishing: 46-47.

Economides, A. & Roupas, C. (2007).Evaluation of Computer Adaptive Testing Systems. In: Int. Journal of Web-Based Learning and Teaching Technologies, 2(1), 70-87, 2007.

European Council (2006). Council Conclusions on the European Indicator of Language Competence. OJ C172, 25/07/2006, p.1, 2006/C172/01, 2006.

European Parliament and the Council of Europe (2006). Recommendations of the European Parliament and the Council on key competences for lifelong learning. OJ L394/10, 30/12/2006. Retrieved 15.5.2008 from http://eur-lex.europa.eu/LexUriServ/site/en/oj/2006/l_394/l_39420061230en00100018.pdf.

Guimares Pereira, A. & Scheuermann, F. (2007). On e-testing: an overview of main issues - Background note. Luxembourg: Office for Official Publications of the European Communities, 2008.

Illes, T., Herrmann, B, Paech, B. & Rückert, J. (2005). Criteria for Software Testing Tool Evaluation – A Task Oriented View. In: Proceedings of the 3rd World Congress of Software Quality, Heidelberg, 2005. Retrieved 15.5.2008 from http://ww.bwcon.de/fileadmin/_primium/downloads/publikationen/IHPR2005.pdf.

Kitchenham, B., Walker, J. (1989). A Quantitative Approach to Monitoring Software Development. Software Engineering Journal, January, 1989

McKenna, C. & Bull, J. (2000). Quality assurance of computer-assisted assessment: practical and strategic issues. In: Quality Assurance in Education, Vol. 8, No. 1, 2000, p.24-31, London.

Pass-it (2005). Guidelines on Online-Assessment. Pass-it project consortium, UK. Retrieved 15.5.08 from http://www.pass-it.org.uk/resources/pass-it_guidelines_on_online_assessment.doc

Plichart, P., R. Jadoul, et al. (2004). TAO, a collaborative distributed computer-based assessment framework built on Semantic Web standards. International Conference on Advances in Intelligent Systems – Theory and Applications; AISTA2004. Luxembourg.

Scalise, K. & Gifford, B. (2006). Computer-Based Assessment in E-Learning: A Framework for Constructing “Intermediate Constraint” Questions and Tasks for Technology Platforms. Journal of Technology, Learning, and Assessment, 4(6). Retrieved 15.05.2008 from http://www.jtla.org.

Valenti, S., Cucchiarelli, A. and Panti, M.(2002). Computer Based Assessment Systems Evaluation via the ISO9126 Quality Model. In: Journal of Information Technology Education, Vol. 1, No. 3, 2002. Retrieved: 15.5.2008 from http://jite.org/documents/Vol1/v1n3p157-175.pdf.

The authors:

Friedrich Scheuermann &
Ângela Guimarães Pereira
European Commission
Joint Research Centre, IPSC,
Quality of Scientific Information
TP-361, Via Enrico Fermi, 2749
I - 21027 Ispra (VA)
E-Mail: Friedrich.scheuermann@jrc.it
Angela.pereira@jrc.it
WWW:
http://kam.jrc.it
http://crell.jrc.it

Friedrich Scheuermann is working for the Centre for Research on Lifelong Learning (CRELL) of the European Commission - Joint Research Centre, IPSC, in Ispra, Italy. He is specialised in the field of technology-enhanced learning and assessment and evaluation research. At CRELL he mainly looks into the role of digital media for skills development and assessment and various aspects of measuring the impact of ICT in education.

Ângela Guimarães Pereira has a PhD in Environmental Engineering and MSc in Sound and Vibration Studies. Leader of the JRC/IPSC Action on Quality of Scientific Information for EU policy making (QSI); she is responsible for activities on science & society interfaces, that range from knowledge quality assurance methodologies to social research and deployment of new information technologies for improvement of European governance of science. For more than a decade she has been designing and implementing ICT based tools to support governance and citizenry dialogue on environmental issues, including (web based and stand-alone) computer applications (e.g. http://kam.jrc.it/vgas; http://b-involved.jrc.it) and mobile applications (e.g. http://mobgas.jrc.it). She is responsible for the network on science and society interfaces: http://kam.jrc.it/ibss. She has also developed guidelines and quality assurance protocols for ICT to be used in public participatory processes.

Computerized Ability Measurement: Some substantive Dos and Don'ts

Oliver Wilhelm & Ulrich Schroeders
Humboldt University Berlin

Abstract
Quality criteria for computerized skills assessment entail general quality criteria that were predominantly established for traditional tests. These traditional quality criteria are not limited to psychometric challenges. The most important challenge that is hard to quantify is the degree to which a test provider succeeds in selecting or deriving a measure for a prespecified purpose. Without profound substantive knowledge this challenge is impossible to accomplish. A second important challenge is the adequate understanding of what an established measurement model represents and what it does not represent. These are the two most important points we want to state. Additionally, we focus on results and implications for cross media equivalence studies and try to derive some perspectives for a research agenda.

Introduction

Much has been said about the desiderata for computer based assessment of skills in other contributions to this volume. We want to focus on a few neglected aspects of these quality criteria. We want to embed our discussion of these aspects into the framework of evidence centered design. The key points we want to make are that a) there are a few pervasive and well known problems of computerized ability measurement that need to be taken serious and b) that there are new still insufficiently exploited opportunities provided by new technologies. We want to make these points on a substantive level rather than on a technological or methodological level.

In principal, administering ability tests through a computer potentially connected to the Internet is not fundamentally distinct from conventional tests. Any ability testing procedure can be characterized in terms of four processes on the assessment delivery layer of the Evidence-Centered Design (Mislevy & Riconscente, 2006; Mislevy, Steinberg, Almond, & Lukas, 2006; Williamson et al., 2004). These four processes are: a) selecting a task, b) presenting the task, c) identifying the evidence collected, and d) scoring the input and providing feedback. We will comment on each of these points before concluding with a brief discussion.

Selection process

In the selection process a task is sampled from a task library or it is generated from a template along with some constraints on the generation. We will discuss this process on two levels of granularity. First, we ask very generally about the nature of the task library from which we sample. Second, we summarize prior experience about the equivalence of tasks across test media.

The nature of the task library
For the present purpose of discussing computerized ability measurement, one important question concerns the nature of the task library. We have put the terms "ability measurement" in the title of this chapter deliberately, thereby deviating from the workshop title which referred to "skill assessment". The reason for this disobedience is that we think that many prevalent labels like "skill" do not allow decent discriminations between constructs. Consider the terms *ability*, *achievement, aptitude, competence, knowledge*, and *skill* for example. In table 1 we list some mainstream definitions of what these terms reflect:

Term	Definition
Ability	– the performance on a cognitive task at present – all mental requirements to fulfill a cognitive task
Achievement	– a result gained by effort – past performance
Aptitude	– like ability and achievement but referring to more specialized abilities in a broader range – the degree of readiness to

	perform well in a particular situation or fixed domain – any characteristic of a person that forecasts his probability of success under a given treatment
Competence	– a context specific cognitive performance disposition that is functionally tied to situations and demands in specific domains
Knowledge	– facts and information acquired through experience or education
Skill	– An ability, usually learned and acquired through training/ practice, to perform actions which achieve a desired outcome – A skill consists of a complex sequence of mental processes, to be performed in a fixed manner

Table 1: Terms for measures of maximal behavior

Now consider your job is to classify existing measures provoking maximal behavior from test-takers. It will be very difficult to come up with dependable classifications of measures or even to name a subset of disjunctive categories for classification. It will also be impossible to derive predictions about the associations of any two measures if you only rely on your or someone else's classifications. Therefore these terms need to be characterized as fuzzy and insufficient when it comes to explaining relations between measures and constructs. The above terms reflect specific research traditions and have no or only little theoretical or empirical substance. Hence, using a single overarching term might be best suited to avoid misunderstandings when referring to measuring maximal behavior, we suggest this term to be labeled "ability". Situations in which maximal behavior is assessed are usually characterized by a) the assessed person being aware of the performance appraisal, b) the assessed person being willing and able to demonstrate maximal performance, and c) the standards for evaluating performance being adequate for assessment (Sackett, Zedeck, & Fogli, 1988).

Within the realm of the so defined task library a variety of distinctions between latent variables is indicated. The most exhaustive effort to date originates from Carroll (1993). Carroll reanalyzed the factor analytic work suggesting a three-stratum theory of human cognitive abilities. Most measures of maximal behavior can easily be classified according to this distinction of mental abilities. In many cases – specifically when it comes to the areas of memory, knowledge, social and emotional abilities – the available evidence is sparse and still insufficient.

Nevertheless, when you get to choose from the task library you should be aware of your degrees of freedom. Unfamiliarity with options in the task library is a serious flaw when it comes to developing a profound understanding of a domain and deriving or using measures from the task-library.

Prior results on equivalence

Meta-analytic studies partly endorse the structural equivalence of ability test data gathered through computerized versus paper and pencil tests (Mead & Drasgow, 1993). For timed power tests the cross-mode correlation corrected for measurement error was r = .97, whereas this coefficient was only r = .72 for speed tests. With reference to the task-library discussed above timed-power tests predominantly represent measures from the domain of reasoning or fluid intelligence, whereas speed tests predominantly reflect clerical speed measures. Neuman and Baydoun (1998) demonstrated that the differences between the two modes can be minimized for clerical speed tests if computerized measures follow the same administration and response procedures as corresponding paper and pencil tests. The authors also tested cross-mode equivalence at three levels that differ in the degree of the restrictiveness of their assumptions about the true scores and errors: parallel, τ-equivalent, and congeneric (Nunnally & Bernstein, 1994). The parallel model is the most restrictive model assuming equal true scores and equal variation of errors. The τ-equivalent model is less restrictive assuming only equal true scores. The congeneric model provided best fit to the data suggesting – in accordance to its restrictions – that tests administered using different media measure the same construct to the same degree, but with different reliability.

In contrast to the relatively large body of literature on psychometric equivalence of computerized and paper and pencil tests (e.g., Clariana & Wallace, 2002; Noyes, Garland, & Robbins, 2004), only a few studies have compared Internet-administrated ability tests to

other test media. In their overview on selection testing, Potosky and Bobko (2004) compared paper and pencil assessments with Internet administration in a simulated selection context. They reported high cross-mode equivalence for an untimed situational judgment test (r = .84) – a procedure difficult to locate in the task-library referred to above and a considerably lower cross-mode correlation for a timed ability test (r = .60), indicating the important influence of time on results in complex tasks. The authors also stated that concerns about equivalence may be more applicable to measures of spatial reasoning or other measures that require visual perception.

Wilhelm, Witthöft, and Größler (1999) tested more than 6,000 subjects on two deductive reasoning tests and an achievement test of business administration knowledge. All three tests were administered both via the Internet and on paper in a between-subjects design without time restrictions. The main difference between the media were higher mean scores for the Internet sample on the reasoning tests, but this difference could easily be attributed to sample characteristics. Subsequent analysis of the data (Wilhelm & McKnight, 2002) with mixture distribution item response theoretical models (Rost, 1997; von Davier & Carstensen, 2007) showed that the administration method had no significant influence on the answer patterns in the ability tests.

Kurz and Evans (2004) reported similar findings for their comparison of a computer- versus a web-based numerical and verbal reasoning test, that is high retest correlations across test media and no significant changes in means and standard deviations. Preckel and Thiemann (2003) compared an Internet versus a paper and pencil version of a figural matrices test for intellectual giftedness. Attributes of the reasoning items contributed in a comparable manner to item difficulty in the online and offline version.

A promising application of testing via the Internet is the field of education. Numerous tests have been developed for the assessment of specific knowledge in order to promote learning. For instance, almost perfect equivalence between a paper-based versus a web-based test in physics could be established (MacIsaac, Cole, Cole, McCullough, & Maxka, 2002).

In sum, these studies support the notion that regardless of test medium and experimental control the same constructs were measured as long as the tests were not speeded. Hence, the not impressive large body of literature supports the following tentative conclusion: If there are no limitations caused by technical constraints, differences between data of an unproctored Internet-testing and a controlled computer-based testing are mainly due to cheating or (self-selected) sample characteristics. Indicators of cheating can be found by detecting and analyzing unexpected responses, that is, correct responses on difficult items in contrast to incorrect answers on easier items, or response time anomalies (van der Linden & van Krimpen-Stoop, 2003). Due to self-selection processes that are difficult to affect Internet test data tend to have slightly higher mean scores. If assessments are used to inform high stakes decisions (e.g., selection tests), there might be stronger differences between an unproctored Internet-based testing and a proctored computer based or paper and pencil based version caused by cheating, a lack of understanding task instructions and the like. Obviously our current knowledge about the equivalence of assessment across test media is by no means sufficient to infer that complex measures can be used regardless of the test medium. It is desirable to clearly distinguish between changes in means and covariances in future studies investigating cross mode comparability. High or even perfect correlations between latent variables of a test administrated in more than one test medium are compatible with substantial changes in means. Therefore, comparisons across test media can privilege participants in one medium over participants in another medium even if the latent variables for the tests are perfectly correlated. Similarly the same test administrated in two test media might have the same mean and dispersion of scores but the two scores might have different reliability and the latent variables captured by both tests might not be perfectly correlated.

Presentation process

In the presentation process a task is presented to the test taker, the interaction of test-taker and items is managed and the results of this interaction are stored.

In theory measures can and should be essentially identical across variations of irrelevant attributes like the test medium selected. However, in practice there is a good chance that you will find differences between a conventional and a corresponding computer-ized test even if a major effort was made to keep tests as equivalent as possible across test media.

However, beside digitizing psychological content and using the computer as a simple electronic page-turner, one of the great advantages of computer-based-testing is the possibility to abandon the static form of a traditional paper-pencil-test and to enrich the material with multi-media extensions or to derive new forms of ability measures that capitalize on the technological opportunities. For example, Vispoel (1999) used the computer's capabilities for audio presentation to develop an adaptive test for assessing music aptitude, or, Olson-Buchanan et al. (1998) compiled short video scenes lasting up to 60 seconds in an assessment of conflict resolution skills. Complex problem solving – supposedly the ability to solve highly complex, intransparent, dynamic, and polytelic problems – were also predominantly measured using computers (Sternberg, 1982; Wittmann & Süß, 1999). In addition to audio and video exten-sions, more and more interactive elements are added. For example, the US Medical Licensing Examination uses interactive vignettes to assess biomedical scientific knowledge and its ap-plication in clinical settings of physicians who are on the cusp of entering medical practice (Melnick, 1990; Melnick & Clauser, 2006). The simulated case studies follow no uniform, rigid course - the setting is interactive instead, thus resulting in an innovative way of assessing medical diagnostic and patient management skills (cp. Kyllonen & Lee, 2005). One important aspect of the above described computer implementations is that the impetus in developing the measures usually was to improve hitherto available forms of measurement. Regularly these and similar new item formats were expected to come closer to reality and to allow assessments with higher externally validity. However, the proponents of ideas that new test media allow for the assessment of "new" constructs so far failed to provide empirical evidence une-quivocally supporting the novelty of latent variables extracted from new measures.

The presentation process is apparently influenced by both software (e.g., layout design, browser functionality) and hardware aspects (e.g., Internet connection, display resolution). Differences between and within test media might therefore not be due to the test medium but to other changes that come along with changing the test medium. For example adaptive versus nonadaptive testing might be responsible for potential differences. However, no differences in test scores obtained by adaptively or conventionally administered computerized tests were found (Mason, Patry, & Bernstein, 2001; Mead & Drasgow, 1993) in the limited empirical evidence available. Another confound that comes along with the change in test medium might be the flexibility of test-taking. In traditional tests it is usually possible to review or skip items and to correct responses. However, there is strong evidence that the majority of examinees will change only a few responses during a review process, usually leading to a slight improvement of performance (Olea, Revuelta, Ximénez, & Abad, 2000; Revuelta, Ximénez, & Olea, 2003).

Some research has been conducted concerning technological questions, for example, the legibility of online texts depending on font characteristics, length and number of lines, and white spacing (for a summary refer to Leeson, 2006). Additionally, the suitability of a specific item/test for computerized presentation needs to be checked specifically unless there are gen-eral rules allowing for a precise prediction of the changes a measure experiences when changing the test medium. Much useful information on presenting item on a computer instead of paper is provided in the International Guidelines on Computer-Based and Internet Delivered Testing (ITC, 2005).

It is also important to realize that using multimedia in testing can also turn into a disadvantage. The opportunity to use multimedia does not imply that using it is always beneficial: Im-plementing audio, video, or animations is time-consuming for test developers and might dis-tract test takers from completing the task. Considering all the pros and cons, we assess that the use of multimedia in ability testing depends on the answer to the question "Is the new medium affecting the measurement process positively?".

No doubt, there are other differences between and within test media, for example, item by item presentation versus itemgroup presentations and it would be worthwhile – albeit labo-rious – to undertake research exploring such potential causes for differences across test media.

Evidence identification process

Work products are assessed by the evidence identification process on the task level. Technically, in ability assessment this is an evaluation according to some performance standard. Such performance standards have veridical character for the traditional assessments of maximal behavior we discuss here.

The possibilities of recording data in computerized measures are manifold, for instance one could easily retrieve mouse pointer movements and mouse clicks, keystrokes, inspection and response latencies or reaction times, IP addresses, and so on.

Nevertheless, collecting more data does not necessarily imply the availability of more valuable information. Therefore, one should consider carefully which behavior to register and which to neglect. For example, relatively little is known about correct decision speed in measuring fluid or crystallized intelligence (Danthiir, Wilhelm, & Schacht, 2005). The critical question is: Would the recording of these data add valuable information to the measurement? For fluid and crystallized intelligence the answer to this question is 'no', if the instruction for a test did not promise credit for fast correct responses. However, for other constructs like mental speed, recording both accuracies and latencies of responses is essential for an adequate assessment of performance. It is important to stress that the possibility to record something does not answer the questions about the utility of this information for diagnostic decisions and the adequacy of this information to satisfy some measurement intention.

In some cases the exact registration of time is critical for substantive reasons. In such cases it can even be critical to compare reaction times across experimental conditions. For example, Linnman, Carlbring, Åhman, Andersson, and Andersson (2006) tested the Stroop paradigm in two versions: a web-administered version and a conventional (offline) computerized imple-mentation. Both versions revealed a strong Stroop effect which also held true for an unproc-tored Internet administration. So, response time measurement in the range of milliseconds is possible via the Internet. If it is necessary to record response times very accurately (e.g., when assessing small effects like negative priming or other indicators of cognitive control), we currently recommend to implement Java-Applets with web-start-technology. Obviously this last recommendation is bound to be valid for a limited time only. We also recommend using proc-tored rather than unproctored test administrations. Proctoring is not only a mean to prevent cheating and other deviant test behavior it also helps to ensure that participants understand and follow the instructions. Other things being equal we would always prefer proctored to unproctored test settings.

It might seem straight forward to handle the evidence identified in ability assessment but there are many methodological problems that are not adequately solved. For example, it is very difficult to distinguish valid from invalid observations. A uni-, bi-, or multivariate outlying datapoint has a strong influence on the means and covariances across observations. Yet these points have a higher probability of indicating invalid observations. These outliers might be persons continuously guessing on a variety of measures or persons who loose motivation half way through a test. In some cases such deviant response patterns might indicate a specific disorder like dyscalculia. Another problem is that the rank order of subjects is unlikely to be the same if we apply different measurement models. Because it is still more the rule than the exception that measurement models are not rigorously tested, there is some ambivalence associated with the person parameter we assign to a specific performance. Sum scores, factor scores or scores from various IRT models will all be highly correlated if a measure isn't seriously flawed, but the score of subjects is not going to be identical. More serious problems arise if more than a single performance indicator is used – like error proportion and number correct in clerical speed tasks or performance on a processing and storage component in dual task working memory measures. Than the integration of two

scores into a single performance indicator is usually warranted and there is a very high proportion of ambiguity associated with such procedures.

A last point we want to raise with respect to the evidence identification process concerns the understanding of scores. We want to discuss this issue in factor analytic terms. In *figure 1* we present various latent variables from various measurement models. The most prevalent case in theory and practice represents measurement per fiat. Here we abstract from an observed vari-able by taking into account its unreliability – as for example expressed in its retest reliability. The disattenuated score is – inadequately – taken as a score for a construct (Borsboom & Mellenbergh, 2002; Schmidt & Hunter, 1999). The model relying on a single latent variable is the case we usually encounter if we establish a fallible measurement model, for example, if we estimate a Birnbaum model or a general factor model based upon dichotomous items. Obviously such a single latent variable will consider task specificity and task class specificity as variance that is due to the latent trait – inadequately so, we surely agree. The method specific latent factor model does not rely exclusively on task A but also includes task B and task C – all using the same method of performance appraisal. Therefore common method variance will be included in the latent factor for this task class. Instead of using raw scores of various measures on the indicator level in this as well as in the next model we could as well establish a lower order measurement model on the task level, that is essentially the single latent variable model discussed before. Finally, in a more general latent factor model we might include a second method of performance appraisal. Here we are interested in performance levels in Fac-tor 12. This might be substantially different from the latent factors discussed before. In fact the latent factor for method 1 is expressed here as a linear function of factor 12 and a nuisance variable indicating method specificity.

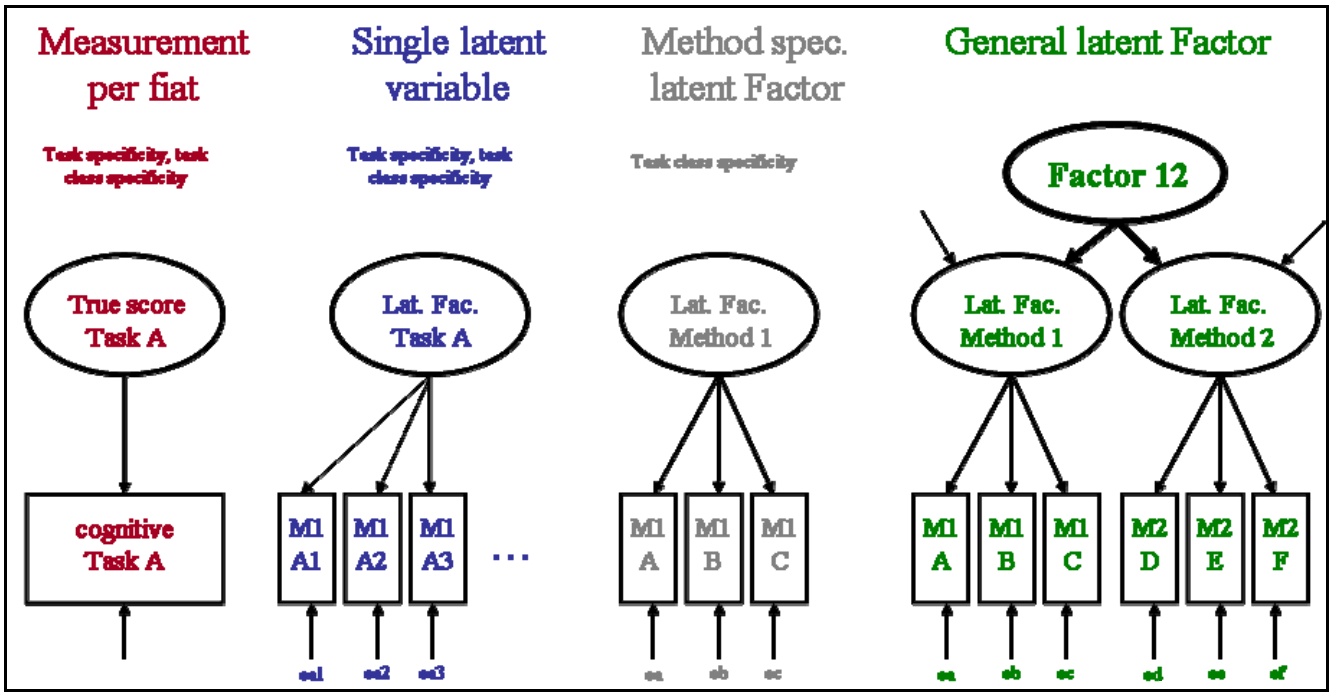


Figure 1: Measurement models of varying methodological sophistication

Obviously it will not always be possible or economical to use a whole battery of indicators in a specific context. However it is always possible to devote some serious reflection to task selection and its implications. The magnitude and relevance of task specificity is a serious and neglected issue when it comes to identify what the available evidence indicates and what it doesn't.

Evidence accumulation and activity selection process

Evidence accumulation or scoring on the level on which psychometric measurement models are usually to be established is discussed here as entailing feedback of test results to the test-taker. Scoring in traditional and computerized measures is essentially identical with respect to the psychometric models applied. A major difference in scoring is the fact that branched and tailored testing is a lot easier to realize in computerized testing. Feedback is likely to be highly similar across test media too but it can be provided immediately through computers - regardless of whether or not the test was administrated through the Internet. Technically, web-based ability testing enables test providers to score item responses directly by accessing an underlying item-database for a specific task from a task library. This database incrementally improves its parameter estimates with every data recording. Regardless of the continuous improvement of parameter estimates, online scoring algorithms can be a very useful thing to implement. A sophisticated scoring algorithm is of interest not only for the selection of the next item, but also for detecting cheating or unwanted item exposure prior to testing.

For the test provider, one great advantage of ability testing via the Internet is that all data gathered from different persons – including those not available for conventional testing sessions, at different locations – including those not accessible with resources usually available, working with different computers – including those usually not at the disposal of the test provider, at different times – including time for which no proctor would be available, can be coded automatically and stored electronically in a central database, ready for analysis. Test materials can easily be altered, updated, or removed during the testing phase. This bliss also is a curse. It is very difficult to tear apart various sources of variance of the test realization that contribute to performance on a test. Therefore high stakes tests completed under different conditions can not be scored by the same algorithm.

Through web-based ability testing the test taker gets the opportunity to receive an automatically generated feedback right after finishing the test. The report can contain a visualization of the testee's results, for example, a distribution of the score in the sample with a marker indicating the subjects position or a template-based text that summarizes the testee's results. Obviously, such reports can be aggregated to the levels of units – like classes or schools. Bartram (1995) annotates that most computer-generated test reports are designed for the test administrator or test provider rather than the test taker. Therefore, one should carefully consider which data to report in what way. Researchers like other test administrators should regard the feedback for the test takers not as a compulsive must but also as an opportunity to communicate individual results as well as the theoretical background of the study and study outcomes. Feedback is not necessarily a one-way street; asking participants for their feedback and opinion can provide valuable information. Hence, apart from the test, a short survey and solicitation for comments can routinely be provided to get test-taker feedback. Apparently the consequences of feedback can be of major interest too. Are test takers adjusting performance relevant behaviors, routines, habits, preferences, values or attitudes as a consequence of receiving a specific feedback? Does the same feedback work the same way in two persons achieving the same results but with diverging personality background? Scoring of test or item results is a pretty straight forward methodological process. In this process there are optimal or at least sound best answers to almost all problems. Feedback of test or item results on the other side seems to be an art much more than a science and much more empirical work is needed in order to adequately capitalize on the technological revolution (Bennett, 2001).

Discussion

In this contribution we attempted to highlight that besides well established general quality criteria there are additional challenges awaiting more intense research efforts in the area of computerized ability measurement. More specifically, intentions to measure constructs like learning ability, civics competence and the like can not be measured by reviewing a few indicators saliently placed on the Internet. What is required is foremost expertise on the subject area. Without profound substantive knowledge measurement intentions that are not substantiated scientifically are doomed to

fail. Dispositions like learning ability or civic competence obviously are primarily or exclusively abilities in the sense they are defined in this chapter. Such dispositions ought to be measured by tests of maximal behavior. We never heard of sport competitions in which athletes were asked about the achievement they were expecting in forthcoming events. How does it occur to social scientists that with ability competitions (as in standardized measures of maximal behavior) self reports of expected, interpolated, or just rated ability levels are just as good as the real thing? We notice such tendencies in the area of social and emotional abilities but also when it comes to learning ability and all sorts of so-called competencies.

We strongly urge test providers to rely on common sense, substantive knowledge, and optimal methodology. Here are some rules of thumb:

- If you don't understand the substantive background of an indicator don't use it without acquiring internal or external expertise
- Make sure you clearly understand the relation between theoretical concepts and derived indicators.
- If you want to measure a construct of maximal behavior, i.e. an ability, use indicators that capture maximal behavior.
- If you can use more than a single measure of an ability use more than a single measure.
- If you use more than a single measure use measures that vary with respect to irrelevant task attributes.
- If you can include covariates in a study use measures that allow you to test convergent and discriminant hypothesis about relations.

References

Bartram, D. (1995). The role of computer-based test interpretation (CBTI) in occupational assessment. International Journal of Selection and Assessment, 3, 31-69.

Bennett, R. E. (2001). How the Internet will help large-scale assessment reinvent itself. Education Policy Analysis Archive [online]. Available: http://epaa.asu.edu/epaa/v9n5.html.

Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. Intelligence, 30, 505-514.

Carroll, J. B. (1993). Human cognitive abilities: A survey of factor-analytic studies. Cambridge, UK: Cambridge University Press.

Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. British Journal of Educational Technology, 33, 593-602.

Danthiir, V., Wilhelm, O., & Schacht, A. (2005). Decision speed in intelligence tasks: Correctly an ability? Psychology Science, 47, 200-229.

International Test Commission (ITC). (2005). International Guidelines on Computer-Based and Internet Delivered Testing [online]. Available: http://www.intestcom.org/guidelines/index.html.

Kurz, R., & Evans, T. (2004). Three generations of on-screen aptitude tests: Equivalence or superiority? In British Psychological Society Occupational Psychology Conference Compendium of Abstracts. Leicester: British Psychological Society.

Kyllonen, P. C., & Lee, S. (2005). Assessing problem solving in context. In O. Wilhelm & R. W. Engle (Eds.), Handbook of understanding and measuring intelligence. (pp. 11-25). London: Sage.

Leeson, H. V. (2006). The mode effect: A literature review of human and technological issues in computerized testing. International Journal of Testing, 6, 1-24.

Linnman, C., Carlbring, P., Åhman, Å., Andersson, H., & Andersson, G. (2006). The Stroop effect on the Internet. Computers in Human Behavior, 22, 448-455.

MacIsaac, D., Cole, R., Cole, D., McCullough, L., & Maxka, J. (2002). Standardized testing in physics via the world wide web. Electronic Journal of Science Education, 6.

Mason, B. J., Patry, M., & Bernstein, D. J. (2001). An examination of the equivalence between non-adaptive computer-based and traditional testing. Journal of Educational Computing Research, 24, 29-39.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. Psychological Bulletin, 114, 449-458.

Melnick, D. E. (1990). Computer-based clinical simulation: State of the art. Evaluation & the Health Professions, 13, 104-120.

Melnick, D. E., & Clauser, B. E. (2006). Computer-based testing for professional licensing and certification of health professionals. In D. Bartram & R. K. Hambleton (Eds.), Computer-based testing and the Internet: Issues and advances. (pp. 163-186). New York: John Wiley & Sons Ltd.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design. In S. M. Downing & T. M. Haladyna (Eds.), Handbook of test development. (pp. 61-90). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Mislevy, R. J., Steinberg, L. S., Almond, R. G., & Lukas, J. F. (2006). Concepts, terminology, and basic models of evidence-centered design. In D. M. Williamson, R. J. Mislevey & I. I. Bejar (Eds.), Automated scoring of complex tasks in computer-based testing. (pp. 15-47). Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Neuman, G., & Baydoun, R. (1998). Computerization of paper-and-pencil tests: When are they equivalent? Applied Psychological Measurement, 22, 71-83.

N**oyes, J., Garland, K., & Robbins, L. (2004).** Paper-based versus computer-based assessment: Is workload another test mode effect? British Journal of Educational Technology, 35, 111-113.

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3rd Ed.). New York: McGraw-Hill.

Olea, J., Revuelta, J., Ximénez, M. C., & Abad, F. J. (2000). Psychometric and psychological effects of review on computerized fixed and adaptive tests. Psicológica International Journal of Methodology and Experimental Psychology, 21, 157-173.

Olson-Buchanan, J. B., Drasgow, F., Moberg, P. J., Mead, A. D., Keenan, P. A., & Donovan, M. A. (1998). Interactive video assessment of conflict resolution skills. Personnel Psychology, 51, 1-24.

Potosky, D., & Bobko, P. (2004). Selection testing via the Internet: Practical considerations and exploratory empirical findings. Personnel Psychology, 57, 1003-1034.

Preckel, F., & Thiemann, H. (2003). Online-versus paper-pencil version of a high potential intelligence test. Swiss Journal of Psychology, 62, 131-138.

Revuelta, J., Ximénez, M. C., & Olea, J. (2003). Psychometric and psychological effects of item selection and review on computerized testing. Educational and Psychological Measurement, 63, 791-808.

Rost, J. (1997). Logistic mixture models. In W. J. van der Linden & R. K. Hambleton (Eds.), Handbook of Modern Item Response Theory (pp. 449-463). New York: Springer.

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. Journal of Applied Psychology, 73, 482.

Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. Intelligence, 27, 183-198.

Sternberg, R. J. (1982). Reasoning, problem solving, and intelligence. In R. J. Sternberg (Ed.), Handbook of human intelligence (pp. 225-307). Cambridge: Cambridge University Press.

van der Linden, W. J., & van Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect aberrant responses in computerized adaptive testing. Psychometrika, 68, 251-265.

Vispoel, W. P. (1999). Creating computerized adaptive tests of musical aptitude: Problems, solutions, and future directions. In F. Drasgow & J. B. Olson-Buchanan (Eds.), Innovations in computerized assessment (pp. 151-176). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

von Davier, M., & Carstensen, C. H. (2007). Multivariate and mixture distribution Rasch models: Extensions and applications. New York: Springer.

Wilhelm, O., & McKnight, P. E. (2002). Ability and achievement testing on the World Wide Web. In B. Batinic, U.-D. Reips & M. Bosnjak (Eds.), Online social sciences. (pp. 151-180). Seattle: Hogrefe & Huber Publishers.

Wilhelm, O., Witthöft, M., & Größler, A. (1999). Comparisons of paper-and-pencil and Internet administrated ability and achievement test. In P. Marquet, A. Mathey, A. Jaillet & E. Nissen (Eds.), Proceedings of IN-TELE 98 (pp. 439-449). Berlin: Peter Lang.

Williamson, D. M., Bauer, M., Steinberg, L. S., Mislevy, R. J., Behrens, J. T., & DeMark, S. F. (2004). Design rationale for a complex performance assessment. International Journal of Testing, 4, 303-332.

Wittmann, W. W., & Süß, H.-M. (1999). Investigating the paths between working memory, intelligence, knowledge, and complex problem-solving performances via brunswik symmetry. In P. L. Ackerman, R. D. P. C. Kyllonen & R. D. Roberts (Eds.), Learning and Individual Differences (pp. 77-108). Washington, DC: American Psychological Association.

The authors:

Oliver Wilhelm
Ulrich Schroeders
Humboldt-Universität zu Berlin
Institut zur Qualitätsentwicklung im Bildungswesen (IQB)
Jägerstraße 10/11
10117 Berlin
Germany

Tel.: +49-(0)30-2093-4873

E-Mail: oliver.wilhelm@rz.hu-berlin.de

Oliver Wilhelm is currently associate professor for educational assessment at Humboldt University. His research interests are focused on individual differences in cognitive abilities.

Ulrich Schroeders is a doctoral student of psychological assessment investigating innovative technologies and their use in the assessment of cognitive abilities.

Challenges for Research in e-Assessment

Jim Ridgway & Sean McCusker
Durham University

Abstract:
A number of research activities are described that are needed to support the development of e-assessment designed to support the progress of the EU. We identify challenges that will shape the design of any e-assessment systems, notably that the goals of the EU are changing rapidly, as is the software environment in which we work. In particular, new software types (such as mashups and folksonomies) serve to redefine our ideas on what is worth knowing, and what is worth being able to do.
We describe research activities under three headings. "Researching the Basics" sets out some obvious targets for research, such as establishing construct validity, ensuring test security, defending against plagiarism and ensuring appropriate access to all users. "Immediate Impact Research" describes important topics that are already the subject of ongoing research that should be explored further, and argues for an ethnographic approach to the uses and impact of new assessment systems. "Impact 'Soon' Research" identifies research topics based on emerging software, and includes ideas such as 'open web' examinations, and a variety of ways that artificial intelligence could be applied. We believe that AI approaches offer ways to solve some difficult assessment challenges.

Pre-amble

In the discussion that follows, we have deliberately chosen a Euro-centric focus. In particular, we address the problems of creating instruments for pan-European surveys in order to monitor the effects of policy, to influence policy, and perhaps to shape policy.

We begin by pointing to some current policy initiatives that require the creation of appropriate tools, and to some disturbing social phenomena that should be reflected in tool design. This paper builds upon an earlier review of e-assessment (Ridgway, McCusker and Pead, 2004) and an update (Ripley, 2007). This earlier content is not repeated here.

The European Context for Research in e-Assessment

The European Union (EU) is at a critical juncture in its development. There are two key drivers of change: one is that the nature of the EU has been changed radically by recent enlargement; the other is that the very definitions of society and societal progress are undergoing reform.

The change in the composition of the EU can be characterized as the addition of relatively poor, but demographically younger countries to relatively rich but demographically older countries. There has been an impressive migration from east to west, in some cases associated with a strain on social services such as education, health and policing. This very rapid migration may well cause problems for community well being. It well known (e.g. Putnam, 2007) that areas of high social mobility are associated with low cultural capital (as measured by indicators such as participation in voluntary work). It is an act of faith to believe that these problems will only be present in the short term, and that long term happiness and prosperity will increase.

In some countries, there is strong evidence for the alienation of some native-born members of minority cultures – examples include riots in Paris suburbs, and the bombing of London tube trains by English Muslims.

This collection of issues is a major driving force behind the need to develop measures of 'intangibles' such as cultural cohesion, alienation, and the like.

A second driving force is a major reconceptualisation of the ways in which the progress of societies should be measured. The current dominant measure is gross domestic product (GDP). Two recent conferences have provided a platform for policy makers such as the president of the EU, and the chief executive of OECD to argue for a much broader range of measures such as cultural capital, happiness of citizens, natural

resources, renewable resources, and infrastructure to become key measures of success. Why is this important? Politicians must be seen to be effective over their term of office, and so are driven by short term pressures to make changes (so that they are seen to be active) and to improve scores on key indicators (e.g. 'waiting times' in the health service). Consider the dilemma of allowing fishing at unsustainable levels. If GDP is the only measure of progress, then, in the short term, over fishing is likely to be allowed. If a broader measure is used that includes renewable resources, then the decision to allow over fishing might lead to a drop in overall societal wealth, and so would be less likely to be taken by politicians seeking re-election. Similarly, UK law now requires that the 1300 companies listed on the Stock Exchange must report on environmental matters and social issues related to their activities (see http://www.tjm.org.uk/corporates/update.shtml).

The Lisbon strategy for growth and jobs (European Commission, 2004) advocates a move towards an economy that is dynamic, competitive, and knowledge based. This requires a workforce that is highly skilled, and able to adapt to new jobs and situations (see Leitch, 2006). There is a perceived need for methods to assess current levels of knowledge concerning 'key skills' or 'key competencies' (KC), and to assess skills of 'learning to learn' (L2L). All of these will be difficult to assess; more problematically, the definition of KC is a moving target. The world of employment is changing very fast, and so definitions of KC will change (as, probably, will notions of L2L).

ICT is a major driver of change. The existence of the web has extended our ideas about the nature of knowledge. Skills in finding information, and critiquing the quality of that information have become more important (for example, the CIA and the Vatican have both been identified by Wikipedia Scanner as editors of some of the changes made to Wikipedia – see http://news.bbc.co.uk/1/hi/technology/6947532.stm).

Recent software developments such as wikis, forums, Facebook, and Many Eyes, are characterised by the construction of knowledge by a community of people rather than by a few individuals. This is sometimes (confusingly) referred to as 'Web 2.0' software. Here, we will use the term 'People Net' (PN) software. Much conventional testing is individualistic and competitive – collaboration is discouraged. PN software has important implications for e-assessment, especially in the context of the definitions of L2L, learning skills, or lifelong learning. PN software may well become an important source for evidence about attainment, and could provide key platforms for testing.

The obvious route to documenting KC and L2L is some form of e-portfolio (the evidence on the effectiveness of e-portfolios to promote learning is, at best, very weak). While some States (e.g. Wales www.careerswales.com/progressfile) have provision for all citizens to maintain a portfolio on a central site, this is not true (so far) across Europe. It is likely that monitoring the state of an individual's KC and L2L will require access to users' own computers. This will present a challenging task, given the plethora of hardware and software systems that are in common use. Monitoring the competence of populations and subpopulations might be handled in other ways (see below).

A symptom of alienation from the education system is the disturbing statistic that 1 in 6 students leave the education system with no qualification (COM 2007, p8). This poses a major challenge for all forms of assessment. Research evidence (Harlen and Deakin Crick, 2002) shows that regular summative assessment has a demotivating effect on low attaining students, and results in poorer learning; it is reasonable to expect that low attaining students will disengage from any assessment system put in place.

A further challenge is the skills base of the communities who educate – professors in higher education, teachers of teachers, and classroom teachers. These groups are not easy to monitor or to influence. Their personal skills, and their ability to foster learning in others of new KC and L2L will be critical to skill and knowledge development, and so must be monitored.

Uncomfortable 'Truths'

It is easy to overstate our chances of success in developing new measures that are ICT based, and that work with our target audiences. As a note of caution, some uncomfortable 'truths' are set out below. Their truth status is ambiguous.

- Working with ICT across the educational sector is particularly difficult, because of the wide range of hardware and software platforms that are used;
- ICT has had very little impact on classroom practices – let alone on attainment;
- Optimistic claims for the likely effectiveness of e-assessment [especially e-portfolio work] are rarely grounded in evidence; such evidence as we have about the benefits of e-portfolios is weak, and discouraging;
- We know far too little about how to design assessment to support learning;
- Change depends on coordinated action across lots of levels in the social system – from political will, through organisational structures, to the actions of individuals – this is very difficult to orchestrate;
- Too many test items are boring!;
- Tests for international or national surveys should not look much like tests for individuals.

Implications for Research Methods

To summarise the discussion so far; research on the practical uses of e-assessment is problematic for a number of reasons:

- System goals (i.e. EU goals) are in a state of flux;
- New artifacts and opportunities will continue to emerge, which will redefine the sorts of knowledge that is valuable;
- The group of students who are the most important to monitor (low attaining and absentee) are likely to be the most difficult to engage in the assessment process;
- There are major organisational barriers to any large scale innovation;
- There are major technical barriers to any large scale innovation;
- We are working in a field where there is over optimism, and too little reality checking.

There are clear implications for research methods. In particular, we need to learn how to:

- Design assessment systems that are sufficiently adaptable to work in a fast changing world;
- Find ways to assess citizens who are disengaged from the learning process;
- Design assessment systems that are robust when working on a variety of platforms and operating systems;
- Understand interactions between layers in the system, and develop methods to monitor and influence on-going processes;
- Expect 'dilution and corruption' and find ways to fix it;
- Learn how to develop high quality items.

Some Research Activities

Learning from On-going Attempts to Introduce Large-Scale Assessment

In England, the government e-strategy document Harnessing Technology (Department for Children, Schools and Families 2005) set out the intention to provide 'online resources, tracking and assessment that works across all sectors, communities and relevant public and private organisations'. Part of this strategy was to introduce some e-assessment into large scale, high-stakes tests, and a great deal has been achieved (e.g. GOLA in the business sector from City and Guilds; the Scottish Qualifications Authority). E-assessment has not received much prominence in recent policy documents. It makes sense to monitor these developments and to share evidence on successes and failures, and to map out guidelines for good practice (e.g. http://www.sqa.org.uk/sqa/files_ccc/guide_to_best_practice.pdf).

Consensus Building and Validation

A key starting point when developing a new measure is to explore different conceptions of the measure, and to think about how any measure might be validated against external criteria. It is sensible to look at existing measures, invent plausible measures, and to explore the psychometric properties of items and subtests.

In language testing, it is clear how one might proceed. It is sensible to review existing tests, and the existing literature on language learning. Tests have blueprints and evidence supporting their factor structure than can be compared, and perhaps synthesized. It is clear how one might validate measures against external criteria – ability to work in the target language in a variety of ways (understanding TV news, ratings of colleagues at work who are native speakers, and the like).

In contrast, L2L, KC, Lifelong Learning and Civics skills all pose big conceptual challenges. The constructs are not clear, nor are the external criteria for validating measures. One might begin with meta-analyses of literature reviews, and content analyses of key policy documents. Focus group discussions conducted in different member states on the nature of the concept, the identification of behaviours that do and do not characterise the concept, and the identification of people who exhibit and do not exhibit the construct will show the extent to which the same words have the same meaning in different countries, and about the possibility of operationalising the construct (Sternberg (2004) has examples of quite different interpretations of 'intelligence' in different countries). Repertory grid techniques are probably appropriate, here (Kelly, 1955).

It may be appropriate to develop core measures that can be applied across the EU, complemented by measures local to countries or regions that are aligned closely with local goals.

An interesting research activity would be to focus on the constructs of groups who are judged to be most disadvantaged in different societies. A comparison of their constructs with the constructs of advantaged groups will give an insight into the scale of the measurement problem we face. The decision about which group's views are adopted is a political one – but one with profound implications.

We offer one conjecture on the likely outcome from such studies – essentially, we would expect to find evidence for a hierarchy of needs of the sort described by Maslow (1943) – with primary needs for food and safety as the major goals for the most disadvantaged groups, and secondary needs such as respect and integration into civic society as the major goals for advantaged groups. If a hierarchical scheme of needs is discovered, there are important implications for both the psychometric approaches that are taken to test development (notably to use techniques suited to discovering and developing appropriate hierarchical scales, such as Rasch scaling) and to reporting – perhaps by the design of a display well suited to representing hierarchically ordered data.

Researching the Basics

Establishing Construct Validity

The section on Consensus Building and Validation pointed to the need for eliciting and operationalising key ideas, and establishing basic information about the nature of the constructs (e.g. the appropriateness of linear and additive models of performance versus hierarchical models), as the basis for a detailed psychometric investigation. Clearly, it is important to explore the psychometric properties of any tests developed:

- To explore the extent to which it is consistent with the test blueprint;
- To look for differences in construct validity between different social groups;
- To look for differences in construct validity between different countries.

In addition to the problems of construct validity discussed above, issues of construct representation will need particular attention if adaptive tests are used (Glas, 2008, this volume), and if test systems provide hints during testing, or ask supplementary questions, when respondents provide partial or incorrect answers.

Backwash

Backwash – or 'consequential validity' (Messick, 1995) or 'generative validity' (Ridgway and Passey, 1993) refers to the changes that occur in any system when a particular high-stakes assessment system is introduced to measure performance. There is an effort to maximize performance on the new indicator and a problem can arise when ways are found to improve scores on the indicator without changing the referent in any way. The 'Texas Miracle' provides an example where dramatic gains were shown on the State test in mathematics, with no associated gains on the National Assessment of Educational Progress

(Klein et al. 2000), which supposedly assessed the same competencies. Exploring backwash effects should be part of the e-assessment research agenda.

One can identify a number of approaches:

- Exploring the TGV (theoretical generative validity – Ridgway and Passey, 1993) via Delphi techniques using a number of different groups;
- Asking stakeholders in different roles to develop plans to 'fake good' on the new indicators;
- Systematic tracking of effects over the short and long term.

In the case of e-assessment, key features to observe will be the motivation and engagement of learners. A surprising recent result from the PISA study where three countries (Iceland, Denmark, and Korea) administered the science component via ICT were the very large differences in student liking for ICT administration. In Denmark and Korea, about 40% of students strongly agreed that they liked ICT administration; in Iceland the comparable figure was just 4% (Björnsson, this volume). These marked differences may well be reflected in future student motivation and engagement with e-assessment.

Technical Issues

The issues of system-wide e-assessment have been addressed (more or less successfully) in a number of countries (e.g. England, Iceland). Sharing information on experiences of successful and unsuccessful implementations will be of great value to the community as a whole.

Security

Security potentially is a major problem for e-assessment. Open source software that is widely used, perhaps paradoxically, offers a degree of security because large numbers of people are motivated to keep their testing systems secure.

For any e-assessment system, there needs to be an on-going development and research programme devoted to anticipating and preventing security breaches. Appropriate strategies include:

- Conformity with established e-assessment standards;
- Real time creation of parallel items (e.g. on a statistics test);
- Establishing a database of problems experienced in the past – attempted breaches, and effective responses;
- Paying hackers to try to breach the system (e.g. the Black Hat group);
- Looking for abnormal patterns of use by testees at the time of testing.

This important issue is dealt with elsewhere in this report.

Plagiarism

Plagiarism is a major problem across the education system (e.g. Underwood, 2006).

People differ a good deal in their definition of plagiarism, and in their ratings of the seriousness of different cheating behaviours (Smith and Ridgway, 2006). Definitions of plagiarism are likely to shift in the light of PN software, where the whole point of the activity is to share and build on others' knowledge, and where the concept of 'authorship' sits uneasily.

Plagiarism in e-portfolios is likely to be particularly problematic, especially if these are used for professional certification, or professional advancement. Some existing techniques are likely to be useful, such as plagiarism checkers (e.g.Turnitin). Style checkers (of the sort used for linguistic analysis) might help, in some contexts, to explore the extent to which different pieces of work in a portfolio were created by the same person.

Work described by Bartram (2008, this volume) provides an excellent base on which to build – for example, looking for items for sale on e-bay, essay sites, and in relevant chat rooms.

Issues associated with the authentication of respondents will be important of the purpose of the testing is to certify (say) L2L. The whole debate on the practicality and morality of biometric systems is relevant here.

It is likely that research in this area will need to continue into the indefinite future. New sorts of fraudulent behaviour will continue to be developed, and the only appropriate response is a programme of research devoted to understanding new mechanisms for cheating, and finding ways to overcome them.

Access

Access is a particular problem for e-assessment. A number of solutions have been developed to facilitate access for paper-based assessment, and a parallel set of provisions needs to be made for e-assessment. As well as obvious factors associated with the properties of the display, such as font size and colour, and background colour (here, choice might actually make access better for some dyslexic users), there are issues related to previous experience with ICT. In some areas of performance (notably when assessing ICT competence), these differences will actually be the focal topic of interest; in others they will be a source of irrelevant variance. If users are actually allowed to choose some features of the display for themselves, this could become a source of error variance if naïve users choose settings ill-suited to their needs.

A critical research issue for pan-European testing relates to access by groups with limited access to technology in the home. Such groups might include economically deprived groups, or cultural subgroups where use of the internet is discouraged. If e-assessment is used to determine access to education or employment, it might exacerbate the problems that disadvantaged groups are currently facing.

Empirical analyses of the relationship between different modes of test administration - perhaps including interviews in the respondents' first language, as an appropriate base line - will be essential to understand instrument effects associated with e-assessment for groups with special educational needs, and for groups who may be disadvantaged by social, cultural and economic factors.

Low Attaining Groups

A group of particular importance to the EU is those people with very low educational attainment. About 1 in 6 students leave formal education with no qualification. It cannot be the case that all these students have no attainments that can be documented and used the basis for further development. This suggests that some simple performance measures be developed that document some basic skills. The Learning and Skills Council (2005) has developed an approach known as Recognising and Recording Progress and Achievement in non-accredited learning, that could be built upon.

Perhaps more significantly, research on the performance of low attaining students (see Harlen and Deakin Crick, 2002) shows that as a result of repeated testing, such students actually disengage from the educational process, and will not attempt to solve problems that they were able to solve earlier in their educational careers. This provides a real challenge for the measurement of lifelong learning, and learning to learn skills. A disengagement from the educational process may well be a sign of cultural alienation, and evidence of poor civic awareness or engagement. If this conjecture is correct, then these students (as students and later as citizens) are unlikely to take up any form of direct assessment, such as e-assessment, with the result that population estimates of competence and civic engagement will be artificially high, and outbreaks of social unrest will come as a surprise to policy makers.

There is a need to conduct research directly into the effects of e-assessment (and other forms of assessment) on low attaining groups – in particular, to explore the negative effects of repeated testing, and to look for ways to design assessment systems that lead to positive responses from low attaining students, i.e. that encourage engagement in the learning process.

Immediate Impact Research

We can identify a number of areas of on-going research that are likely to have an immediate impact. Some of these are set out below.

Exploring the ethnography of e-assessment

Pan-European e-assessment is a completely new venture. There is an urgent need for an ethnographic approach – studying the activity patterns of different stakeholders, from students to policy makers, in terms of their actions, and the implications of their actions, for the education of individuals and the design of large-scale education systems.

In principle, there will be rich opportunities for self assessment, self diagnosis, and an active engagement with rich learning resources. This brave new world could promote the development of engaged, autonomous learners with well developed L2L skills. The actual impact of access to rich assessment might be rather different from this optimistic view.

Everyone a psychologist?

There is a plethora of models that advocate constructivist approaches to learning and assessment, and some sites that facilitate this sort of assessment (e.g. ALTA http://www.alta-systems.co.uk/demos.html#AMS; mCLASS Reading) A key set of questions concern the circumstances in which these approaches are effective. Examples include:

- Mapping learning pathways;
- Exemplifying goals;
- Offering learning strategies;
- Supporting and stimulating reflection;
- Encouraging peer evaluation.

These are often associated with the idea that adult learning can be different from children's learning in important ways. In particular, because adults have had many previously successful learning experiences, their L2L skills can be used to further future learning. We have a great deal to learn about adult learning that will be valuable in the design of e-assessment, especially for L2L and lifelong learning.

Automated processing of free text input

Research on the automated processing of free text input ranges from genuine semantic analysis of short paragraphs (notably in science (e.g. Sukkarieh et al., 2003) and medicine (e.g. Mitchell et al., 2003), through latent semantic analysis (e.g. Landauer, 2002 – implemented by Pearson Learning as the Intelligent Essay Assessor™) to essay marking on the basis of surface features of text such as the number of keystrokes. All of these approaches can have positive educational benefits; these need to be explored in detail.

Assessing process skills in creative areas

Kimbell's e-scape project (http://www.teru.org.uk/) sets out to assess process skills associated with design. Students record interim results via PDAs, drawing sketches, taking photographs, and recording their reflections. The assessment is done in a structured way, and results are managed electronically. Studies on the consistency of grading between judges are very positive (e.g. inter rater reliability of 0.93 – see Kimbell, 2007). This work provides evidence that complex process skills can be assessed reliably, and in a way that is resistant to plagiarism. Important research could be conducted, exploring and validating similar approaches in other creative areas.

Reducing assessment by moderating teacher assessments

The assessment burden on students and teachers may well be problematic for students and teachers (Harlen and Deakin Crick, 2002). Research designed to reduce external assessment via a greater use of teacher judgments moderated electronically should be explored. Teacher ratings could also be evaluated by survey methods on short tests to schools.

Using large scale surveys to guide policy

Large scale surveys can identify lacunae in student knowledge. There can be a tighter relationship between assessment and policy. For example, in the World Class Tests, studies that identified specific weaknesses in the performances of high-attaining students were used to design curriculum materials for all students, focussed on these weaknesses.

E-Portfolios

E-portfolios offer an obvious approach to the challenge of documenting life long learning, KC and L2L skills. The research literature on e-portfolios is rather sparse, and the claims made for the likely benefits of e-portfolios far exceed any evidence we have about the successful use of e-portfolios. There is an urgent need for appropriate research. The provision of lifelong 'learning spaces' to anyone in Wales who wants such a facility should be monitored carefully. This project seems to be well designed. There are good reasons for citizens to take advantage of the resources provided – such as advice on the preparation of curriculum vitae, tools for self diagnosis and reflection, career ideas (including support for people being made redundant, and for retirement), information on job vacancies, and the like.

Providing students with descriptions of the domain of study may well be valuable – mapping out developmental stages, and exemplifying key target behaviours, could well be beneficial to learners.

The efficacy of different forms of e-portfolios needs to be evaluated.

Impact 'Soon' Research

More use of conventional web applications
A number of uses of conventional web applications are set out below. These will all need extensive development and research before they are viable approaches to large scale testing.

'Open web' examinations seem eminently sensible – many problems faced by professional people are approached via collaboration between physically close and remote colleagues, and by extensive use of the web. The ability to use the web effectively is an important KC, and learning how to use new tools is an important L2L skill. Search strategies during open web examinations may well provide useful summative information on current skill levels, and formative information to guide future learning.

Some professions require regular professional training, and the periodic demonstration of competence as part of professional licencing. Validation of qualifications could be mediated via RSS. Participants would be given a limited amount of time to respond to tasks that would demonstrate their competence.

On demand testing can be developed further. For example, ,the eVIVA project allows students to book a test on line, then to answer pre-recorded questions by phone, that are scored by human markers after a short time delay.

Use of 'People Net' Web Resources
A number of web resources are being created that facilitate collaboration between people, where users can upload content, collaborate actively, and where the expertise derives from the whole community, rather than from a few experts. Here we will call these resources 'People Net' (PN) resources (rather than 'Web 2.0' resources). PN resources, potentially, are important for assessment for a number of reasons, and we have barely begun to explore their potential. PN resources allow different sorts of performance - notably more authentic performance - to be assessed. Facility with PN tools is an emerging KS. Some possible uses are set out below. In this section, we focus on using PN resources with the full knowledge and consent of participants. The nature of the activities are clear, and any assessment systems could be described in such a way that participants could judge (and improve) their performance. We discuss covert monitoring in the next section.

Mashups such as popfly, netvibes and pipes allow users to combine data from multiple sources into a single tool, so survey data can be overlaid onto Google maps, for example. Mashup editors can accommodate RSS feeds. Mashups could be submitted as evidence of KS, or as evidence of substantive domain knowledge in a particular area;

Wikis such as wikipedia can provide evidence on procedural skills associated with collaborative writing and organizing information (as well as demonstrating skills in finding information). Student contributions to wikis and forums could be assessed using tools such as wiki scanner; their ability to contribute, and to learn from such resources could be a component of any attempt to assess their L2L capability. It is easy to imagine an extension of wikis where users are invited to test their declarative knowledge of the sections of the wiki that they have just studied;

Wiki 'skeletons' (e.g. http://www.wiki.com/) could be used with individuals or small groups to support the creation of knowledge in a key domain. These creations could then be assessed in terms of dimensions such as semantic structure, completeness, and sense of audience (judged in terms of ease of navigation and use of language, for example);

Assessment wikis could provide space for expert contributions of particularly useful individual tasks that can then be used freely - this could have a very positive impact on the culture of teaching and learning; gathering student votes on the most useful tests or test items (in the style of Amazon) should be explored;

Folksonomies (or collaborative tagging or social bookmarking) such as del.icio.us and Furl are methods of describing content in terms of user-defined tags, in ways that can be shared. The ways in which sources are tagged, and the sources that are identified, provide insights into the user's semantic constructions. Use of others' tagging to identify resources quickly is an indicator of KC. Taxonomic tasks could be devised to make such assessments more formal;
Blogs and e-portfolios, and any form of diary keeping and recording information, can support reflection, and can demonstrate the authenticity of work by tracking the development of ideas and products;

Communication tools such as Discussion Forums, Skype and MSN, facilitate interviews and information exchange, and provide evidence for authentication. Youtube offers the facility to upload video, and can show evidence of the ability to create and share information (for example, by uploading a series of web pages assembled via clipmarks). These could be used to demonstrate the authenticity of student work, or as a substantive demonstration of learning in a particular domain;

Search engines such as Google can provide information on user skills in finding information;

Social networking tools such as Facebook can provide evidence on networking skills that could be assessed as a component of citizenship. They could also provide a portal where student contributions to blogs, discussions and the like, are assembled using a unique student identifier.

There are interesting developments on workplace uses of PN for professional development (e.g. Brown et al., 2007). These activities include the use of e-portfolios to support reflection, and the use of PN to build a workplace community that innovates, that supports professional development, and that supports organisational change.

Artificial Intelligence Approaches
Artificial Intelligence (AI) approaches have been used to address a wide range of problems where sophisticated pattern recognition is likely to be useful. Here we outline some applications to assessment. In particular, we identify ways in which estimates might be made about ICT competence, KC, L2L and about a number of aspects of citizenship, alienation, social and cultural capital, and the like, in ways that are non-obtrusive. Unobtrusive measures have a number of potential benefits. There are few problems with scale – larger numbers of participants do not necessarily require much more processing – indeed, large numbers of participants are needed to develop appropriate categorisations. Biases associated with verbal self reports will be avoided – as will the problems of differential drop out by members of different ethnic (or alienated) groups. AI systems are easily extensible – new items do not have to be written and piloted, as would be the case with measures of new educational goals developed via conventional psychometric methods.

Embedded assessment
Embedded assessment has many advocates (e.g. Birenbaum et al., 2006). One conception of embedded assessment is that students engage in learning activities and in performative tasks as part of the normal pattern of learning and instruction, and that an assessment system draws conclusions about their competencies based on what they actually do. For example, recognising competencies in ICT would be relatively easy to do, given access to all the keystrokes on someone's personal computer. A large number of approaches can be taken to the description of performances and to the recognition of performances (including Bayesian models, simple pattern recognition systems and AI connectionist networks). More obvious forms of assessment – such as testing students formally, when the AI system judges they will pass easily – could be used to complement this approach.

Survey data – engagement with PN
Some PN sites (e.g. Facebook) provide information on users, categorized in a variety of ways. These data could be useful in identifying overall levels of engagement in PN activities, and (more interestingly) in identifying the patterns of activity in vulnerable subgroups. AI approaches could be used to categorise individuals in terms of their likely sex, ethnicity and social group.

AI analyses of traffic on data nodes could provide valuable information about patterns of ICT use, broken down by region, ethnicity and sex, and social background. Essentially, the software could be based on software designed to detect terrorist activities, or on the data mining techniques used for focused marketing by supermarkets, applied to 'loyalty card' transactions.

Process skills assessed via e.g. Google desktop and spyware

This is an extension of the concept of embedded assessment. Spyware would be loaded (with user knowledge) onto users' computers, and would track the patterns of activity. Here the topic under investigation is the way that the student searches for information, the sources used, and the like. Given an appropriate research design (c.f. PISA sampling of respondents) such a system could be used to collect large scale survey data on a wide range of competencies.

This data analysis could be used at the individual level to provide feedback to the user to improve their search strategies, and to identify areas for their future development – for example, by categorising contributions to discussion forums using a variety of analytic schemes, such as De Bono's (1985) 'Six Thinking Hats', or Bales' (1950) Interaction Process Analysis.

Review

There are a number of challenges to the development of e-assessment designed to support the activities of the EU. Some of these challenges are obvious (such as the problems of large-scale innovations in complex environments, technical and security issues, and the challenges of engaging with groups of citizens that are hard to reach); some are less obvious and harder to plan for, such as the changing goals of the EU, and the emergence of new software artifacts that actually serve to redefine our ideas on what is worth knowing, and being able to do. The latter requires a period of consensus building on what is important to assess, and ways to design assessment systems that are 'future proofed'.

Here, we described some research agendas under a number of headings. Researching the Basics addressed key psychometric topics such as establishing construct validity, security, plagiarism and access. Immediate Impact Research set out research activities that included an ethnographic approach to the uses and impact of new assessment systems, a critical examination of the impact of systems that encourage user autonomy, and further research on systems designed to assess creative skills, and to automate the processing of free text. Impact 'Soon' Research identified topics for research that include exploring the potential of 'open web' examinations, the use of 'people net' resources such as mashups, folksonomies and wikis, and a variety of ways that artificial intelligence could be applied.

We believe that AI approaches have considerable potential for providing evidence on some of the most difficult assessment challenges we now face. In particular, large scale assessment of process skills (including specific ICT skills, L2L and KC), and of cultural integration, disaggregated by region, ethnicity, sex and conventional academic attainment become possible. These measures are likely to be more authentic than verbal self reports. They require little or no direct attention from citizens, and so are likely to be far less vulnerable to sampling bias than are conventional tests.

References

Bales, R. (1950). A set of categories for the analysis of small group interaction. American Sociological Review, 15, pp. 257 – 263.

Bartram (2008) this volume.

Birenbaum, M., Breuer, K., Cascallar, E., Dochy, F., Dori, Y., Ridgway, J., Wiesemes, R. and Nickmans, G. (2006). A Learning Integrated Assessment System, Educational Research Review, 1(1), pp 61-67.

Bjornsson (2008) this volume

Brown, A., Bimrose, J., and Barnes, S-A. (2007). Personalised Learning Environments, Portfolios, and Formative Assessment in the Workplace. http://www.assessnet.org.uk/file.php?file=/1/Resources/Conference_2007/Alan_Brown_paper.pdf

Commission of the European Communities (2007). The European Interest: Succeeding in the age of globalisation. COM(2007) 581 final.

De Bono, E. (1985). Six Thinking Hats. Harmondsworth: Penguin.

Department for Children, Schools and Families (2005). The e-Strategy – Harnessing Technology: transforming learning and children's services. www.dfes.gov.uk/publications/e-strategy

Effective Practice http://www.sqa.org.uk/sqa/files_ccc/guide_to_best_practice.pdfoint Information Systems Committee (2006) e-Assessment: an overview of JISC activities. http://www.jisc.ac.uk/uploaded_documents/ACFC6B.pdf

European Commission (2004). Facing the Challenge: The Lisboa Strategy for Growth and Employment, Brussels. http://ec.europa.eu/education/policies/2010/doc/kok_en.pdf

Glas, C. (2008) this volume.

Harlen, W., and Deakin Crick, R. (2002). A Systematic Review of the Impact of Summative Assessment and Tests on Students' Motivation for Learning. London: EPPI Centre. http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=108

Kelly, G.A. (1955). The Psychology of Personal Constructs. Norton: New York.

Kimbell, R. (2007): http://www.goldsmiths.ac.uk/teru/UserFiles/File/e-scape2.pdf

Klein. S.P., Hamilton, L.S., McCaffrey, D.F., and Stecher, B.M. (2000) What Do Test Scores in Texas Tell Us? RAND Issues Paper. http://www.rand.org/publications/IP/IP202/

Landauer, T. K. (2002). On the computational basis of learning and cognition: Arguments from LSA. In N. Ross (Ed.), The Psychology of Learning and Motivation, 41, 43-84. http://lsa.colorado.edu/

Learning and Skills Council (2005). Recognising and recording progress and achievement in non-accreditied learning. http://readingroom.lsc.gov.uk/lsc/2005/quality/performanceachievement/recognising-recording-progress-achievement-july-2005.pdf

Leitch, S. (2006) Prosperity for all in the global economy – world class skills. http://www.hm-treasury.gov.uk/media/6/4/leitch_finalreport051206.pdf

Maslow, A. (1943). A Theory of Human Motivation. Psychological Review, 50, 370-396. http://psychclassics.yorku.ca/Maslow/motivation.htm

Messick, S. (1995) Validity of Psychological Assessment. American Psychologist. Vol. 50. No. 9. Pages 741-749.

Mitchell, T., Aldridge, N., Williamson, W., and Broomhead, P. (2003). Computer based testing of medical knowledge. Proceedings of the 7th International Computer Assisted Assessment Conference, Loughborough, pp249-267.

Putnam, R. (2007). E Pluribus Unum: Diversity and Community in the Twenty-first Century. The 2006 Johan Skytte Prize Lecture. Scandinavian Political Studies, 30(2), 137-174.

Ridgway, J., McCusker, S., and Pead, D. (2004). Literature Review of E-assessment. NESTA Futurelab: Bristol. pp. 48. http://www.nestafuturelab.org/research/reviews/10_01.htm

Ridgway, J., and Passey, D. (1993). An International View of Mathematics Assessment - through a glass, darkly. In M. Niss (ed.). Investigations into Assessment in Mathematics Education. Kluwer: London. pp. 55-72.

Ripley, M. (2007). E-assessment – an update on research, policy and practice. Bristol: Futurelab. http://www.futurelab.org.uk/resources/publications_reports_articles/literature_reviews/Literature_Review204

Smith, H., and Ridgway, J. (2006). Another Piece in the Cheating Jigsaw. Second International Plagiarism Conference, JISC, Newcastle, UK.

Sternberg, R. (2004). Culture and Intelligence. American Psychologist, 59(5), 325-338.

Sukkarieh, J., Pulman, S., and Raikes, N. (2003). Auto-marking: using computationa linguistics to score short, free text responses. Paper presented to the 29th Annual Conference of the International Association for Educational Assessment. http://www.aqa.org.uk/support/iaea/

Underwood, J. (2006). Digital Technologies and Dishonesty in Examinations and Tests. http://www.qca.org.uk/libraryAssets/media/qca-06_digital-technologies-dishonesty-exams-tests-report.pdf

Web-Sites:

ALTA: http://www.alta-systems.co.uk/demos.html#AMS

Black Hat: http://www.blackhat.com/

City and Guilds: http://www.cityandguilds.com/cps/rde/xchg/SID-0AC0478D-0BFB4FAA/cgonline/hs.xsl/660.html?vroot=14

Clipmarks: http://clipmarks.com/

del.icio.us: http://del.icio.us/

eVIVA: http://www.qca.org.uk/qca_5962.aspx

e-scape: http://www.teru.org.uk/

Facebook: http://www.facebook.com/

Furl: http://www.furl.net/

Google: http://www.google.co.uk/

GOLA: http://www.goalonline.co.uk/Web/goalonline/

Intelligent Essay Assessor™: http://www.knowledge-technologies.com/

Many Eyes: http://services.alphaworks.ibm.com/manyeyes/home

mCLASS Reading: http://www.wirelessgeneration.com/products.php?prod=mClass:Reading3D

MSN: http://www.msn.co.uk/

Netvibes: http://www.netvibes.com/

Pipes: http://pipes.yahoo.com/pipes/

Popfly: http://www.popfly.ms/

Scottish Qualifications Authority: http://www.sqa.org.uk/sqa/5606.html

Skype: http://www.skype.com/

Turnitin: http://turnitin.com/static/index.html

Youtube: http://www.youtube.com/

Wales: www.careerswales.com/progressfile

Wiki.com: http://www.wiki.com/

Wikipedia: http://en.wikipedia.org/wiki/Main_Page

Wikipedia Scanner: http://news.bbc.co.uk/1/hi/technology/6947532.stm

World Class Tests: http://www.worldclassarena.org/

The authors:

Jim Ridgway and Sean McCusker
Durham University
School of Education, Leazes Road, Durham DH1 1TA, UK
www.dur.ac.uk/smart.centre/

E-Mail: jim.ridgway@durham.ac.uk and
E-Mail: sean.mccusker@durham.ac.uk

Jim Ridgway is Professor of Education and Sean McCusker is a Research Associate at Durham. Both have research interests that include: assessment and e-assessment; data visualisation and understanding; mathematics education, including gender issues and applicable mathematics.

Important Considerations in e-Assessment

An educational measurement perspective on identifying items for an European research Agenda

Gerben van Lent
ETS Global BV

Abstract
In 2005 The International Test Commission published the International Guidelines on Computer Based and Internet Delivered Testing. The Guidelines are divided in four sections: technological issues, quality issues, levels of control; and security and safeguarding privacy. In this paper two aspects related to quality issues are highlighted because of their crucial importance in developing fair, valid and reliable tests: comparability of scores and adaptive testing. The paper looks into research that has been conducted regarding comparability, and adaptive testing to date and formulates eight broad areas that could be of interest for a European Research Agenda. It concludes that the future of CBA might lie in continuous testing when there is a clear need but fixed-dates otherwise, testing on flexible locations instead of in test centres and test development for, rather than translated to computerized administration.

Introduction

In preparing this paper I 'googled' the internet and selected the first test I came across about creative problem solving. I took the test by answering the questions (36 in total) randomly and solved two problems by again randomly selecting an answer. Within 10 seconds I received a summary report that looked more or less like this:

Snapshot Report

Openness to Creativity

0 **55** **100**

Your responses indicate that you're a "planning problem-solver". Essentially, although you won't let your imagination run wild, your approach to problems does strike a good balance between creativity and practicality. Although you may prefer to base your strategies on logical rules and conventions, you also try to think "outside the box". People who score similarly to you are both *reproductive* and *productive* thinkers. Depending on the problem at hand, you'll either use similar problems encountered in the past as a reference (reproductive), or look at it from different perspectives and search for alternative ways of solving it (productive). Overall, this is a fairly good approach to use. Remember however, that while there's generally nothing wrong with taking a more practical approach to problems, it does impose some limitations. By taking that step outside your standard way of thinking and expanding your imagination, you'll not only be able to increase your options but may end up uncovering ideas that had never crossed your mind before!

Obviously my behavior violated some of the assumptions the test producers must have had when developing the test. As was indicated this test was developed 'to evaluate whether your attitude towards problem-solving and the manner in which you approach a problem are conducive to creative thinking.

In 2005 The International Test Commission published the International Guidelines on Computer Based and Internet Delivered Testing (http://www.intestcom.org/guidelines/index.html). The Guidelines are divided in four sections: technological issues, quality issues, levels of control; and security and safeguarding privacy.

In this paper I wish to highlight two aspects related to Quality Issues:

- Comparability of Scores
- Adaptive testing

Firstly I will give some background about ETS, the organization I work for, in the context of computer based and internet based testing. The portfolio of tests from ETS that are administered via computers with or without making use of the internet, ranges from some 'indicative' tests linked to language proficiency that are available to everyone, practice tests via a subscription model, high stakes tests related e.g. to language proficiency, critical thinking and problem solving, or subject specific which are delivered in dedicated test centres only or in classrooms. The test themselves include multiple choice, constructed response, or scenario based questions. Stimulus material can be written, oral, video and/or pictures. Scoring can be analytical or holistic and might include process evaluation besides the response

ETS or individual researchers from ETS have participated in many research projects (research papers: http://www.etsemea-customassessments.org/) related to computer based and internet based testing partly to support the development of our own tests so as to meet our Standards for Quality and Fairness (http://www.ets.org/Media/About_ETS/pdf/standards.pdf), partly to contribute to more general research projects either within ETS or with and for external partners.

The goals of the Network are described as to work on a research agenda related to e-assessment and issues which have not been covered yet by research approaches and programs of the European Commission. In the next sections I will suggest possible suggestions for a research agenda and make reference to research that should be conducted for each computer based test separately.

Comparability of Scores

In the fourth edition of Educational Measurement (Brennan 2006) comparability is described as "*the commonality of score meaning across testing conditions… including delivery mode and computer platforms*". If we pretend that scores are comparable, while in fact they are not we may make wrong decisions. It can affect decisions about promotion, graduation, diagnosis, progress reporting, hiring, promotion or training. It would directly violate fairness principles.

Let me highlight a few aspects that are related to comparability.
Especially in large scale programs often, for a while at least, two modes of delivery exist: paper based and computer based. How do we know that the scores obtained through these different modes are comparable and can be used interchangeably or that scores obtained through one mode can be linked to cut scores of another mode? When scores are equivalent they indicate individuals are rank ordered in the same way and the score distribution are approximately the same. If distributions are not the same, methods for equating can facilitate interchangeability. But when is it possible to equate and when should you decide that the two tests are actually measuring something different and cannot or should not be equated?

It is interesting to observe that while in paper based testing we are very concerned about preserving the integrity of items; wording, lay out, position in the test, it seems we are much less concerned when items show up in different ways across platforms, the clarity of pictures, the lay out on the screen, the size and resolution of the screen, etc.
How much do we know of the effects on answering an item correctly by the response requirements? Does it make a difference whether you tick a correct answer on a mark-sheet, or write texts in a paper based environment or click a mouse, drag and drop or write text using a key board. Do spelling and grammar play the same role, how do you treat the use of text messaging codes?

➧ 1: **It seems that a research agenda should as a minimum identify key issues that are related to comparability**

ETS has e.g. conducted or participated in research related to (see references: ETS research reports):

- Comparability of Delivery Modes
- Comparability of Test Platforms.

The first type of comparability with respect to delivery modes, in the context of e-assessment, has to do not only with whether the scores from paper and computer tests mean the same thing but increasingly also relates to desktop, laptop,

personal digital assistants (pda), mobile phone, etc. In the field of large scale assessments, the research has been conducted mostly with respect to the equivalence of scores from paper based tests and computer based tests. This includes research about differences resulting from presentation characteristics, response requirements, general administration characteristics, timing and speediness characteristics.

In the fourth edition of Educational Measurement (Brennan 2006, p. 502) results of research in this field in the US is described. Results seem to indicate that while they all show that there are differences; in most cases they are not significant. Not many large scale studies have been conducted in primary and secondary education but there are a few linked to the National Assessment of Educational progress (NAEP) in the US). They showed that at least at the time of the research (2001-2002) computer familiarity showed up as a significant source of irrelevant variance in online test scores. One might hope that with the increased penetration of computers in society these effects will diminish over time.
Nevertheless once a paper and pencil test is transformed into an online test, it is always necessary to conduct research about the comparability of the two tests.

➧ 2: **It could be advisable to identify guidelines at a European level for the research that has to be conducted to show that a paper based and its equivalent online tests are indeed comparable.**

The recent tender of the European Commission regarding the development of a European Indicator for Language Proficiency where for the time being two delivery modes would be included is a good illustration about the necessity of such guidelines.
Our observations included amongst others the following suggestion that research be an integral part of the total operation:

Comparability of Administration Modes
The use of two methods of administration, a computer delivered and a paper and pencil delivered assessment, will likely introduce mode effects that need to be taken into account when analyzing and comparing results. The computer delivered assessment may be advantageous for some tasks, while the paper and pencil assessment may be so for other types of tasks. Familiarity with technology may vary across countries and groups of students, and so may the familiarity of students with paper and pencil based tests.

The possible effects limit the comparability of results across modes of administration. In order to account for these undesirable effects, comparability studies will need to be conducted. These studies would require the cooperation of countries that request the availability of the computer-based delivery mode before that option can be implemented.

The comparability or "bridging" studies would coincide with … cycle of the main survey data collection. Within that data collection, sub-samples of randomly equivalent test-takers within the participating countries would be assigned to either a paper and pencil or a computer-based test. The overall comparability of the administration modes would be evaluated through test level analyses (similar to equipercentile equating). These analyses will produce the adjustment values necessary to convert the computer-based scale scores onto the paper and pencil based scale. Item level analyses will point to the type of items that may be more susceptible to mode effect. [Excerpt from ETS Global BV response to Restricted call for tender n° EAC 21/2007: "European survey on language competences"]

The second type of comparability with respect to test platforms is related to the differences in software and/or hardware that cause different forms of item presentation. This includes differences in internet connectivity, screen size, screen resolution, operating systems settings and browser settings. They can lead to differences in font size on the screen, amount of text on the screen, amount of scrolling. Research conducted for the SAT in the US indicates that only scrolling had a significant effect. Measures to be taken to promote comparability include setting standards for equipment to be used and to make use of software that 'takes over' the candidate's computer for the test. In large scale e-assessments it is very important that the circumstances of the test taker don't differ across candidates.

⧫ 3: In the design of comparability studies it would be important to consider upfront what data are needed for a meaningful study instead of conducting research on the basis of data that happen to be available.

Test Design Considerations

In e-assessment the concept of Adaptive Testing is very popular, because it seems such an elegant way of gathering information: each candidate is tested at his/her own level and a relatively small number of items per candidate seems to be needed. When people talk about adaptive testing they often use the term for anything that is not a fixed test form given to all students. So before highlighting a number of issues I will first introduce a brief taxonomy of test designs with some indication of advantages related to conventional tests (Davey & van Lent, 2007):

Fixed Forms
These are conventional tests delivered by computer and constructed like conventional tests. All examinees see all items and the tests are scored like conventional tests. They are as efficient as and as secure as conventional tests. Multiple test forms can be developed and used interchangeably if properly equated.

Random Forms
Tests are drawn from an item pool. They are constructed on-line following specified rules and each examinee sees only a portion of the items available. They can be scored like conventional tests (but not as robustly) or by previously IRT-calibrated items.

They are as efficient as conventional tests and more secure than conventional tests.

Semi-Adaptive Tests
These are also known as stratified or multi-stage. Examinees are routed through a series of pre-assembled item blocks. Routes taken are dictated by performance. Tests can be scored conventionally (by equating) or by previous IRT calibration. They are more efficient than conventional tests and more secure than conventional tests. They are well-suited to naturally set-based items.

Classification Tests
These tests have the goal of classifying examinees, rather than producing a numeric score (i.e., examinees are simply sorted into groups, such as pass or fail). Students respond to items until a criterion for a pass/fail decision is met. Each examinee sees only some of the items available and the items ideally are targeted at decision threshold. Test length is meant to vary across examinees and ideally tests are scored by decision theoretic methods.
Because numeric scores are not produced, classification tests can be very efficient and they can be more secure than conventional tests.

Adaptive Tests
With conventional tests, most questions are too easy or too hard for most examinees. We learn little from asking these questions. Much more is learned by asking questions for which answers are least predictable. The goal of a Computer Adaptive Testing is to identify these questions and ask them.
By tailoring the test to the examinee, a CAT can be both short and precise.
Work best from a fair-sized pool of available items and the items must have known properties (obtained via pretesting). Constructed Response items pose problems.

Hybrid Designs
These designs combine several test administration strategies in order to achieve multiple goals. E.g. a combination of multi-stage and classification testing would be a compelling choice for a situation where both normative and formative (diagnostic) information are to be gathered from a single test.

A Multi-stage / Classification Hybrid
This consists of a broad-range multi-stage test that surveys the entire substantive domain and efficiently produces a single score accurately positioning a student on the general construct.
Each item would also contribute to one or more diagnostic sub-scores that are graded on a two- or three-category proficiency scale (e.g. master / non-master or basic/proficient/advanced).
Completing the multi-stage test may already allow classification decisions in some sub-score domains to be reliably made. Classification testing would continue in remaining domains until all decisions are made.

Issues with forms of non linear testing (Schaeffer et al. 1995)

There are still many issues to be addressed regarding large-scale adaptive or other non fixed testing programs.

⧫ 4: It seems that a research agenda at a minimum should focus on key aspects around adaptive testing resulting in transparency what conditions should be met and what procedures should be followed to build a non fixed testing program.

I will highlight a number of specific issues.

What is an optimal configuration of item pools related to level of exposure and stakes of the test?
In fixed form testing programs, one set of questions is administered to large numbers of persons on a single day. Thus, item exposure is limited to a short period of time. Item exposure is determined by the security measures around test administration and how many times the test form will be used.
In forms of adaptive testing this becomes more complicated. Compared to a fixed test that is used more than once exposure may be lessened, because there is no guarantee about the items a candidate will get other than that they come from the same pool. However the size of the pool and the frequency of testing will have a big influence on item exposure and thereby test security
If CAT pools are to be in operation for long periods of time, this level of exposure would become commonplace.

How can the quality of a pool be built, monitored and maintained over time?
A key design issue relates to how to monitor item quality over time in adaptive tests. It is common practice that item parameters will be calculated based on pretests with a sample with wide range of abilities. However when they are used they will be administered to individuals with a narrow range of ability that is directly linked to the level of difficulty of the item. Are we then still 'talking' about the same item or has it changed its characteristics? Another issue is that the characteristics of the items are determined as they are included in the item pool. Afterwards the assumption is made that these characteristics remain stable, regardless of exposure of the particular item or for that matter the family of items to which that item belongs. In other words are item characteristics stable independently of exposure?

⧫ 5: It seems that a research agenda could also include identifying methods to establish equivalence of item parameters derived from pretesting in a traditional setting and from adaptive and considering new methods of evaluating pretest data.

What is needed to assure equivalence of e-assessment in international settings?
Although research conducted to date has demonstrated that adaptive and traditional tests can be comparable, the increased use of computer adaptive testing throughout Europe raises issues. Research like what was done for the Graduate Records Examinations® indicate that people with little or no computer familiarity can learn the testing system and use it effectively in a short period of time.
However, whereas in many countries candidates might be quite familiar with technology it can never be just assumed that this is not a factor that impacts on the ability of a candidate to do the test.

⧫ 6: It seems that a research agenda could also include an inventarisation of current practices regarding the assumptions about computer familiarity when tests are made available.

Will adaptive testing result in differences in traditional patterns of differences among subgroups?
The results of various comparability studies demonstrate that we can achieve comparability of non fixed and adaptive tests. However not so much research has been conducted whether this also holds for subgroups. You could hypothesize that tests that are targeted at a specific difficulty level will decrease frustration and therefore improve performance on the test and if this would affect subgroups more strongly using adaptive tests could be considered an improvement. In Europe if tests are meant to be used across countries this might be particularly relevant.

⧫7: It seems that a research agenda could include developing best practices on scrutinizing performances of subgroups.

What opportunities and problems do computer adaptive tests create with regard to testing individuals with special education needs?
In Europe 'inclusion' is seen as very important. The potential of the computer to provide alternatives to traditional test modifications is apparent. You can think of alternative input devices, recording tests, sign language on screen, modification of screen displays (larger

fonts, colour), recording responses instead of writing, just to name a few. But should we consider this as simple modifications or do they change the test in a significant way. Should these modifications be accessible to all test takers or only specific modifications to designated groups? If you receive a certificate or a formal score, should it be indicated on the score report or certificate that you were tested with a modified administration mode?

⧫ 8: It seems that a research agenda could include investigating the tools that are available to support computer based testing for candidates with special needs and how they impact on test administration, performance and the concepts of standardized testing.

Conclusion

Currently many of the changes in computer based testing seem to be driven first and foremost by what technology allows us to do. Educational research tries to catch up, but is lagging behind. For major high stakes testing programs this often leads to a situation that relatively 'old' methods of testing are used: either paper based tests or computer based tests that are more or less paper based tests put on screen. However some innovation is included in the form of new closed items formats, multi media stimulus material and adaptive testing formats.

The test users are often frustrated at this perceived lack of flexibility or turn to more innovative test providers only to be frustrated later on by the lack of reliability and validity of the scores that are produced.

A good example are computer based tests that indicate they also measure process by tracking the activities students undertake on the computer while answering a question or performing a task. Experiences of ETS (www.ets.org/iskills)] and of the Qualification and Curriculum Authority (QCA) (Key Stage 3 ICT test see www.naa.org.uk/naaks3) indicate that developing these tests is far from trivial both in aspects of linking the recorded activities of the candidates to credible models of proficiency and in developing tasks that have parallel demands.

The current state of Computer Based Testing can maybe be characterized by the table below:

The Current State of CBT	Good Fit Cases
• Some CBTs offer little or no value added. • Some "innovative" items are likely to contribute more 'artifactual' than valid measurement. • Other items are starting to exploit capabilities. • Limited site capacity often forces continuous administration, which can introduce serious security concerns. • Test administration algorithms are getting smarter but remain limited.	• Practice exams. • Formative / diagnostic assessments. • Placement tests. • Small volume / high-stakes certification. • Technical certification.

When looking to the future one would hope that the following characteristics would be accurate descriptions:

- Tests will be administered continuously only if there is good reason to do so.
- Tests will be administered at "sites of convenience" rather than dedicated test centres.
- Tests will be developed for, rather than translated to computerized administration.
- Test administration will be whole lot "smarter."

Key issues that have to be taken care of in realizing this state are: Design Issues, Accessibility Issues and Security Issues. Appropriate underpinning with relevant research is an absolute necessity.

For a European Research Agenda I have identified in this paper eight areas mainly related to Comparability and Adaptive testing:

1. To identify key issues that are related to comparability
2. To identify guidelines at a European level for the research that has to be conducted to show that a paper based and its equivalent online test are indeed comparable
3. To ensure in the design of comparability studies upfront what data are needed for a meaningful study instead of conducting research on the basis of data that happen to be available.
4. To focus on key aspects around adaptive testing resulting in transparency what conditions should be met and what procedures should be followed to build a non fixed testing program
5. To identify methods to establish equivalence of item parameters derived from pretesting in a traditional setting and from adaptive settings and considering new methods of evaluating pretest data.
6. To conduct an inventarisation of current practices regarding the assumptions about computer familiarity when tests are made available
7. To develop best practices on scrutinizing performances of subgroups
8. To investigate the tools that are available to support computer based testing for candidates with special needs and how they impact on test administration, performance and the concepts of standardized testing.

References:

Brennan, R. 2006 (Ed.). 4th edition of Educational Measurement, jointly sponsored by the American Council on Education (ACE) and the National Council on Measurement in Education), summer 2006.

Davey, T. & van Lent, G. (2007). Practical Considerations in Computerized Testing for AQE (Association for Quality Education) Northern Ireland, December 2006.

Schaffer, G.A., Steffen, M., Golub-Smith, M.L., Mills, C. N. & Durso, R. (1995). The Introduction and Comparability of the Computer Adaptive GRE General Test, GRE Board Report No. 88-08aP, August 1995.

ETS research reports:

Schaeffer, G.A., Bridgeman, B., Golub-Smith, M.L., Lewis, C., Potenza, M.T., and Steffen, M. (1998). Compara-bility of Paper-and-Pencil and Computer Adaptive Test Scores on the GRE® General Test (Research Rep. No. 98-38). Princeton, NJ: Educational Testing Service

Schaeffer, G. A., Steffen, M., Golub-Smith, M. L., Mills, C. N., and Durso, R. (1995). The Introduction and Comparability of the Computer-Adaptive GRE® General Test (Research Rep. No. 95-20). Princeton, NJ: Educational Testing Service.

Sandene, B., Horkay, N., Bennett, R. E., Allen, N., Braswell, J., Kaplan, B., & Oranje, A. (Eds.). (2005). Online assessment in mathematics and writing: Reports from the NAEP Technology-Based Assessment Project (NCES 2005-457). Washington, DC: National Centre for Education Statistics, US Department of Education.

Bridgeman, B., Lennon, M.L., and Jackenthal, A. (2003). Effects of Screen Size, Screen Resolution, and Display Rate on Computer-Based Test Performance. Applied Measurement in Education, 16(3): 191-205.

The author:

Gerben van Lent
ETS Global BV
Strawinskylaan 913
1077XX Amsterdam
The Netherlands

E-Mail : gvanlent@etsglobalbv.org

Gerben van Lent (1957) has been Executive Director of ETS Global BV since October 2006. He leads the Custom Assessment Solutions Unit of ETS Global BV worldwide. He Represent the organisation by offering advice on assessment and learning solutions to clients and prospects directly or in support of ETS Global BV country managers. He represents ETS Global BV at conferences and seminars and leads pilot and key projects. He supports the managing director in coordinating strategic initiatives and is responsible for the strategic, organisational and operational development of custom assessment solutions activities in the region. He has contributed articles to journals in the field of education and measurement and he is an external member of the Research Committee of Assessment and Qualifications Alliance (AQA) in the United Kingdom and is Board member of the Arabic Centre for Educational Testing Services in Jordan. He joined ETS in August 2001. Before his appointment as Executive Director, Gerben van Lent was Director of Business Development of ETS Europe from 2004 to 2006, with similar responsibilities but limited to the European region.

Prior to his role at ETS, Gerben van Lent was Senior Advisor of the International Department of CITO (National Institute of Educational Measurement) in the Netherlands. He was responsible for business development and project management, leading amongst others 3 education reform projects in Eastern Europe and developing policy papers for education reform.

European Commission

EUR 23306 EN – Joint Research Centre – Institute for the Protection and Security of the Citizen
Title: TOWARDS A RESEARCH AGENDA ON COMPUTER-BASED ASSESSMENT - Challenges and needs for European Educational Measurement
Editors: Friedrich Scheuermann & Angela Guimarães Pereira
Luxembourg: Office for Official Publications of the European Communities
2008 – 106 pp.
EUR – Scientific and Technical Research series – ISSN 1018-5593

Abstract
In 2006 the European Parliament and the Council of Europe have passed recommendations on key competences for lifelong learning and the use of a common reference tool to observe and promote progress in terms of the achievement of goals formulated in ?Lisbon strategy? in March 2000 (revised in 2006, see http://ec.europa.eu/growthandjobs/) and its follow-up declarations. For those areas which are not already covered by existing measurements (foreign languages and learning-to-learn skills), indicators for the identification of such skills are now needed, as well as effective instruments for carrying out large-scale assessments in Europe. In this context it is hoped that electronic testing could improve the effectiveness of the needed assessments, i.e. to improve identification of skills, by reducing costs of the whole operation (financial efforts, human resources etc.). The European Commission is asked to assist Member States to define the organisational and resource implications for them of the construction and administration of tests, including looking into the possibility of adopting e-testing as the means to administer the tests. In addition to traditional testing approaches carried out in a paper-pencil mode, there are a variety of aspects needed to be taken into account when computer-based testing is deployed, such as software quality, secure delivery, if Internet-based: reliable network capacities, support, maintenance, software costs for development and test delivery, including licences. Future European surveys are going to introduce new ways of assessing student achievements. Tests can be calibrated to the specific competence level of each student and become more stimulating, going much further than it can be achieved with traditional multiple choice questions. Simulations provide better means of contextualising skills to real life situations and providing a more complete picture of the actual competence to be assessed. However, a variety of challenges require more research into the barriers posed by the use of technologies, e.g. in terms of computer, performance and security. The “Quality of Scientific Information” Action (QSI) and the Centre for Research on Lifelong Learning (CRELL) are carrying out a research project on quality criteria of Open Source skills assessment tools. Two workshops were carried out in previous years bringing together European key experts from assessment research and practice in order to identify and discuss quality criteria relevant for carrying out large-scale assessments at a European level. This report reflects the contributions made on experiences and key challenges for European skills assessment.

The mission of the JRC is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.